DOCUMENT RESUME

ED 254 162 HE 018 127

TITLE	Statistical Sampling Handbook for Student Aid Programs: A Reference for Non-Statisticians. Winter 1984.
INSTITUTION	Office of Student Financial Assistance (ED), Washington, DC.
PUB DATE	84
NOTE	111p.; For related documents, see HE 018 112-135 and HE 018 137-140.
PUB TYPE	Guides - Non-Classroom Use (055) -- Tests/Evaluation Instruments (160)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Administrator Guides; College Students; Computation; *Federal Aid; Government Employees; Higher Education; *Prediction; Program Administration; *Records (Forms); Sample Size; *Sampling; *Statistical Analysis; *Student Financial Aid; Student Financial Aid Officers
IDENTIFIERS	*Office of Student Financial Assistance

ABSTRACT

A manual on sampling is presented to assist audit and program reviewers, project officers, managers, and program specialists of the U.S. Office of Student Financial Assistance (OSFA). For each of the following types of samples, definitions and examples are provided, along with information on advantages and disadvantages: simple random sampling, stratified sampling, cluster sampling, systematic (interval) sampling, dollar-unit sampling, sequential (stop or go) sampling, discovery (exploratory) sampling, multi-stage sampling, opportunity sampling, and quota sampling. Forms to aid in calculating a variety of common sample statistics are included. Three examples of the potential uses of sampling statistics and the forms by OSFA are provided, and potential problems that could arise are addressed. The forms are used to: calculate the estimate of the population variance from a sample; develop population estimates from a simple random sample; determine minimum necessary sample sizes; illustrate the use of a calculator to determine population variance; and develop population estimates. Appendices include: an introduction to the mathematics of sampling, information on sampling formulas and symbols, a 13-item annotated bibliography, and an index by primary reference or definition. (SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Office of Student Financial Assistance / U.S. Department of Education

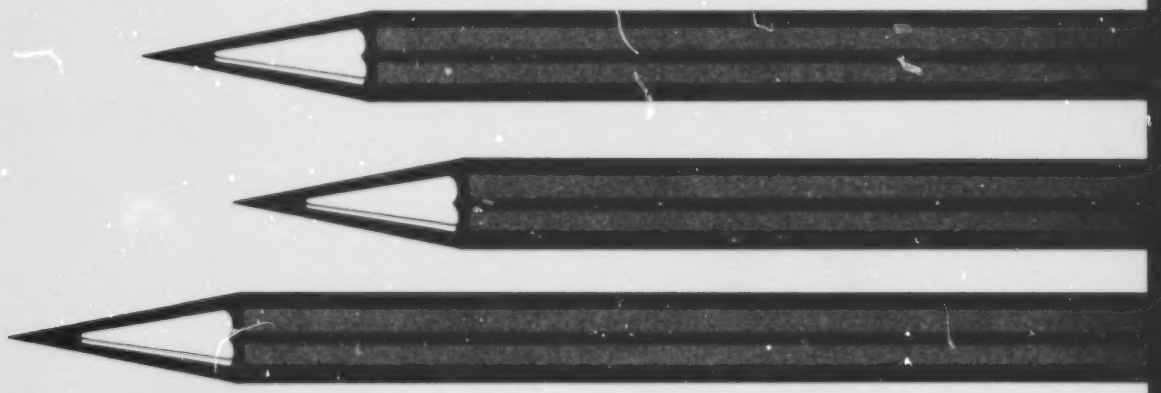
ED254162

SCOPE OF INTEREST NOTICE
The Eric Facility has assigned this document for proccesing to: HE
In our judgment, this document is also of interest to the Clearinghouses noted to the right. Indexing should reflect their special points of view TM

Winter 1984

Statistical Sampling Handbook for Student Aid Programs

A Reference for Non-Statisticians



U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

HE 18 127

STATISTICAL SAMPLING HANDBOOK

FOR STUDENT AID PROGRAMS

A Reference for Non-Statisticians

Prepared by

Division of Quality Assurance

Office of Student Financial Assistance

Office of Postsecondary Education

U.S. Department of Education

WINTER 1984

TABLE OF CONTENTS

Chapter		Page
1	WHEN TO SAMPLE	
	Introduction	1
	Using the Manual	2
	The Advantages and Disadvantages of Sampling	3
	Judgmental and Statistical Sampling	5
2	THE LANGUAGE OF STATISTICAL SAMPLING	
	Introduction	9
	Population and Sample	10
	Sampling Error	10
	Point and Interval Estimates	13
	Confidence Intervals, Confidence Levels, and Confidence Limits	13
	Summary	14
3	TYPES OF SAMPLES	
	Introduction	15
	Simple Random Sampling	16
	Using a random number table to draw a simple random sample	17
	Stratified Sampling	21
	Cluster Sampling	22
	Systematic (Interval) Sampling	23
	Drawing a systematic sample from filing cabinets	24
	Dollar-Unit Sampling	26
	Sequential (Stop or Go) Sampling	28
	Discovery (Exploratory) Sampling	29
	Multi-Stage Sampling	31
	Opportunity Sampling	32
	Quota Sampling	33
	Summary	34

TABLE OF CONTENTS (Continued)

Chapter		Page
4.	COMPUTING SAMPLE STATISTICS	
	Introduction	35
	Form A: Estimating the Variance of a Population from a Sample	36
	Form B: Developing Population Estimates from a Simple Random Sample	38
	Form C: Determining Sample Sizes	41
	Form D: Using a Calculator to Compute the Estimate of the Population Variance from a Sample	47
	Form E: Using a Calculator to Compute Population Estimates from a Simple Random Sample	48
5.	APPLICATIONS OF SAMPLING TO STUDENT FINANCIAL AID	
	Introduction	51
	Example 1: Review of SEOG Awards at University A	52
	Example 2: BEOG Applicant Quality Control Study	60
	Example 3: Review of CWS Audit Report from University B	61
APPENDIX A:	INTRODUCTION TO SAMPLING STATISTICS	
	Introduction	A1
	Total, Population Size, Mean, Variance, Standard Deviation and Distribution	A2
	Attributes	A7
	Population and Sample Symbols	A9
	Estimating the Population Mean	A11
	Estimating the Population Total and Standard Deviation	A12
	Standard Error of the Mean	A13
	Confidence Interval for the Mean	A15

TABLE OF CONTENTS (Continued)

Chapter		Page
	Small Samples	A17
	Confidence Interval for the Total	A18
	The Relation Between the Confidence Interval and the Confidence Level	A20
	The Relation Between Sample Size and Population Size	A22
	Self-Testing Review	A25
	Answers to Self-Testing Review	A27
APPENDIX B:	SAMPLING SYMBOLS AND FORMULAS	
APPENDIX C:	BIBLIOGRAPHY	
	INDEX	

1

When To Sample

For many research and audit purposes, it is not practical or possible to collect data on all the cases in the population under study but only on some fractional part called a sample. A sample is chosen to represent the total population from which it is drawn. If properly constructed, sample characteristics can provide the basis for making valid statistical inferences about the total population.

Sampling is not the sole province of statisticians, but occurs in hundreds of ways in everyday life. The cook samples soup to correct the seasoning. The elementary school teacher quizzes students on a sample of arithmetic problems to determine the students' overall mathematical abilities. The political pollster samples the electorate to predict the outcomes of elections.

Virtually every division of the Office of Student Financial Assistance employs data based on statistical samples. However, three groups within OSFA are particularly frequent users of sample data. These three groups are audit and program reviewers in the Division of Certification and Program Review (DCPR) and project officers, managers and program specialists in a variety of divisions.

The DCPR audit reviewers are recipients of sampling based audit data which they review, analyze and summarize for reports. In addition, they are often called upon to defend their understanding of the data when recipient institutions challenge audit findings. These tasks require the ability to develop population estimates from sample data and a basic understanding of statistical sampling and its application to program auditing.

Each year DCPR conducts hundreds of program reviews at recipient institutions. During the course of the reviews, a selection of cases at each institution are examined for compliance with program procedures. Although, in general, formal statistical procedures are not employed in selecting cases for examination, program reviewers could benefit from a basic knowledge of statistical sampling techniques. Such knowledge would permit reviewers to employ statistical sampling procedures where appropriate as well as to advise institutions on methods of conducting internal reviews.

Project officers, managers, and program specialists comprise the third group of frequent sample data users. The sampling needs of members of this group are less specialized and more general than those of the other two groups. Common tasks requiring knowledge of statistical sampling include: designing samples for quality checks; selecting samples of students and/or institutions for research studies; and reviewing sampling plans and sample data submitted by contractors.

USING THE MANUAL

This manual is designed to address the statistical sampling needs of the groups listed above and OSFA as a whole. Because almost all potential users of this manual are not statisticians, mathematical

exposition and technical language have been kept to a minimum. The manual concentrates on explaining, in clear and nonmathematical language, the issues raised in sampling, the utility of sample data, and methods of calculating required sample sizes and making population estimates from sample data. Manual chapters have been written so that, to the maximum extent possible, each chapter stands by itself. This format allows use of the manual both as a reference source and as a self-teaching introduction to sampling.

The manual is divided into two primary sections. The first section, chapters one, two, and three, introduce the advantages and disadvantages of sampling, the terminology of sampling, and the major types of samples. The focus of these chapters is conceptual, not mathematical. The second section, chapters four and five and appendices A, B, and C, introduce the statistics of sampling in several different, and largely independent ways. Chapter 4, Computing Sample Statistics, contains a series of forms to aid in calculating a variety of common sample statistics. Chapter 5, Applications of Sampling to Student Financial Aid, gives examples of statistical sampling uses in OSFA and illustrates the use of the forms contained in chapter 4. Appendix A, Introduction to Sample Statistics, is an optional resource for those who want a basic introduction to the mathematics of sampling. Appendix B summarizes basic sampling formulas and symbols. Finally, for those who would like a fuller explanation of sampling statistics or more advanced or specialized statistical information, Appendix C presents a short annotated bibliography.

THE ADVANTAGES AND DISADVANTAGES OF SAMPLING

In many cases, drawing a statistical sample has a wide range of potential advantages over examination of the entire population under study.

- Reduced costs. For many of the needs of OSFA auditors and program reviewers, it is not practical or possible to examine all the student files in an institution under review. In such cases a sample of student aid files can often produce the

necessary information at a fraction of the cost of a 100% review. Similarly, OSFA researchers can often statistically describe the total population of student financial assistance recipients on the basis of sample data when a complete census of all recipients would be prohibitively expensive.

- Reduced Respondent Burden. Sampling allows for shorter reviews and less disturbance of reviewed agencies than would be possible with 100% reviews.
- Greater accuracy through better quality control. When a smaller number of records is being reviewed, data collection, analysis and summarization may be more carefully supervised and controlled than might be possible in a review of all the documents required in a 100% review.
- Greater range of information obtainable. Sampling permits the researcher, auditor, or program reviewer to examine a wider range of topics than would generally be practical if a 100% review were required for every topic addressed.
- Faster reporting of results. The time required for data collection and summarization can be greatly reduced through the use of sampling.

The use of sampling, however, is not always appropriate, or without difficulties. The primary reasons for not using sampling are:

- Sampling results in incomplete knowledge. By its very nature, sample data can only produce estimates of population characteristics. For example, from a sample of student financial aid records, it is possible to estimate the total dollar amount of student aid overpayments. It is not possible, however, to determine the exact dollar amount of overpayment or which students received overpayments and which did not.
- Sampling introduces potential bias. When a sample is not correctly constructed, the conclusions based on the sample can be biased. For example, the famous Literary Digest poll that predicted Alf Landon would defeat Franklin Roosevelt for president in 1936 reached the wrong conclusion because of a faulty sampling technique. Individuals the sample polled were selected out of telephone directories. In 1936, telephone subscribers were among the more prosperous of the voting population and, as a group, predominantly supported the Republican Landon.
- Sampling produces only aggregate statistics. If information is needed for every individual in the population, sampling is inappropriate.

- Sampling can be inappropriate when studying small populations (Under 100 cases). When the number of files to be reviewed, records to be audited or population to be studied, are small enough it may become logistically and administratively simpler not to sample and to do a complete review.

In summary, sampling can be a useful tool for reducing review, audit and research costs, and respondent burden; for maintaining high quality control; for expanding the scope of research; and for speeding the reporting of results. Sampling, however, is not appropriate when exact knowledge of a population characteristic is required or when information is needed for individual cases in the population. Faulty sampling can produce biased results. Because sampling introduces certain complications in audits, program review and research efforts, its advantages do not always outweigh its disadvantages. This is particularly true when studying small populations.

JUDGMENTAL AND STATISTICAL SAMPLING

Once the decision has been made to sample, it must be decided whether formal statistical sampling procedures should be followed in constructing the sample. There are two primary types of samples: judgmental (or purposive) and statistical (or probability). For a judgmental sample, cases are chosen for study on the basis of the selector's knowledge and experience. For a statistical sample, one or another variation of random selection is employed in choosing cases for study. In Chapter 3 the major types of judgmental and statistical samples are reviewed. Here the more general question of the relative advantages and disadvantages of statistical sampling versus judgmental sampling is discussed.

JUDGMENTAL SAMPLING

Judgmental sampling can be an efficient method of locating cases of interest. For example, a program reviewer may have learned from experience that procedural errors are more likely to be found in thick, dog-eared files than in thin, clean files. Therefore, in a review to discover whether procedural errors exist, selecting only thick dog-eared files may be more efficient than statistical sampling.

Judgmental sampling can be more efficient than statistical sampling in describing a population on the basis of very small samples. Research has shown judgmental sample selection is most effective when the sample is small (eight or less); when the population sampled is small and visible or known to the selector; and when the selector has great and proven skill in this art.

Judgmental samples can involve fewer complications than statistical sampling. Judgmental sampling can eliminate elaborate statistical sampling procedures and analysis.

STATISTICAL SAMPLING

Statistical sampling is superior to judgmental sampling in a number of very important ways. Results of statistical sampling are objective and defensible. Because statistical sampling rests on demonstrable, mathematical principles, the results are objective and defensible before reviewers, recipients, and even courts. Questions of bias or bad judgment which can be raised against judgmental samples can be eliminated through statistical sampling.

Results of statistical sampling provide a sound basis for drawing inferences about the total population from which the sample was drawn. For example, examination of a statistical sample of student aid files in a university could provide the basis for estimating the total number of aid overpayments in that university. In contrast, a judgmental sample of thick or dog-eared files, while it might be efficient in locating particular errors, could not serve as the basis for estimating the total number or size of errors. This is because judgmental samples violate the statistical principles which make possible projections from a sample to a total population.

Statistical sampling provides an estimate of sampling error. For judgmental samples there is no way of knowing whether two different samples are likely to produce the same or widely divergent results. There is also no method of determining how sample results are likely to compare with the results that would be obtained from a 100 percent review of all cases in the population. Statistical sampling, however, produces

estimates of sampling error. For example, if a simple random statistical sample of 200 loan records out of 2,000 student loans made by a single lender found 20 of the loans delinquent, it would be possible to estimate that the chances are 95 in a 100 that the total number of delinquent loans for the lender would be somewhere between 121 and 279. (Chapter 4 presents an explanation of how such estimates are calculated.)

Statistical sampling provides an objective means of determining necessary sample size to meet program review, audit or research purposes. Returning to the previous example, assume that the Federal reviewer determined that it is necessary to estimate the number of delinquent loans for the lender within a range of plus or minus 50 loans. To achieve this level of accuracy, it is possible to determine in advance the necessary minimum sample size of 433 cases.

Statistical sampling results may be combined and evaluated even when conducted in different locations and by different individuals. As an example, results of statistical, sample-based audits, independently conducted by different auditors, on various campuses of a single university system, can easily be combined and analyzed to produce estimates for an entire university system. Because procedures for selecting judgmental samples inevitably vary from auditor to auditor and from circumstance to circumstance the results of judgmental seldom can be combined.

Statistical sampling is flexible enough to incorporate most of the advantages of judgmental sampling. Statistical samples can be tailored to particular needs and circumstances in a great variety of ways which allow incorporation of the auditor's, program reviewer's or researcher's knowledge and experience. If, for example, a program reviewer has advance knowledge that thick files were more likely to contain errors, a stratified statistical sampling method could be developed which gives thick files a higher likelihood of being selected than thin files. Such a sampling method would incorporate the advantages of judgmental sampling while retaining all the advantages of statistical sampling.

The choice between judgmental and statistical sampling must be made on a case by case basis. Judgmental sampling is effective when the sample and the population sampled are very small and visible or well known to the selector and the selector has skill and experience in drawing such samples. Statistical sampling is superior where objective, defensible results are required or where projections to the total population are to be made or where the sample size is twenty-five or larger. In most cases, the advantages of statistical and judgmental sampling can be combined by tailoring statistical samples to particular circumstances and needs.

2

The Language of Statistical Sampling

Statistical sampling has its own special language. The language is composed of common English words which are given special meanings, letters from the Roman and the Greek alphabets and mathematical notation. Although, at first view, this language can be intimidating, the basic underlying concepts are very straightforward. This section introduces the basic language and concepts of sampling.

POPULATION AND SAMPLE

The dictionary's first definition of population is all the people in a country or region. In statistics, the term population is used much more broadly to mean the total set of items, persons, files etc. from which a sample is taken. Items which compose a population could be individual students, receipts, apples, universities, files or any other set of entities to be studied. A sample is any portion of the population selected to be studied.

A sampling unit is a selected item or case from or about which information is sought. It is often possible for the sampling unit to be defined a number of different ways in the same area of study. For example, in an audit of student financial aid, sampling unit could be defined as the individual recipient or as each financial aid award.

A measure which describes a population is called a parameter. For example, if the population under study is a year's BECG awards in a particular university, the number of awards, the total dollar amount of awards, and the average amount of the awards in the university could all be parameters. A statistic is a characteristic of a sample. If we were to draw a sample of BEOG awards in the university, the number of awards in the sample, the total dollar value of the sample awards and the average dollar value of sample awards are all statistics. When we make generalizations about a population on the basis of sample data, we are using statistics to estimate parameters. To help maintain the distinction between parameters and statistics, Greek letters such as σ, and τ are generally used to denote parameters and lowercase Roman letters to denote sample statistics. Table A7 on page A9 summarizes the basic symbols and formulas used in statistical sampling.

SAMPLING ERROR

Estimates of population parameters can be calculated from sampling statistics. By its very nature, sample data can produce only estimates of population parameters. For example, from a sample of student aid

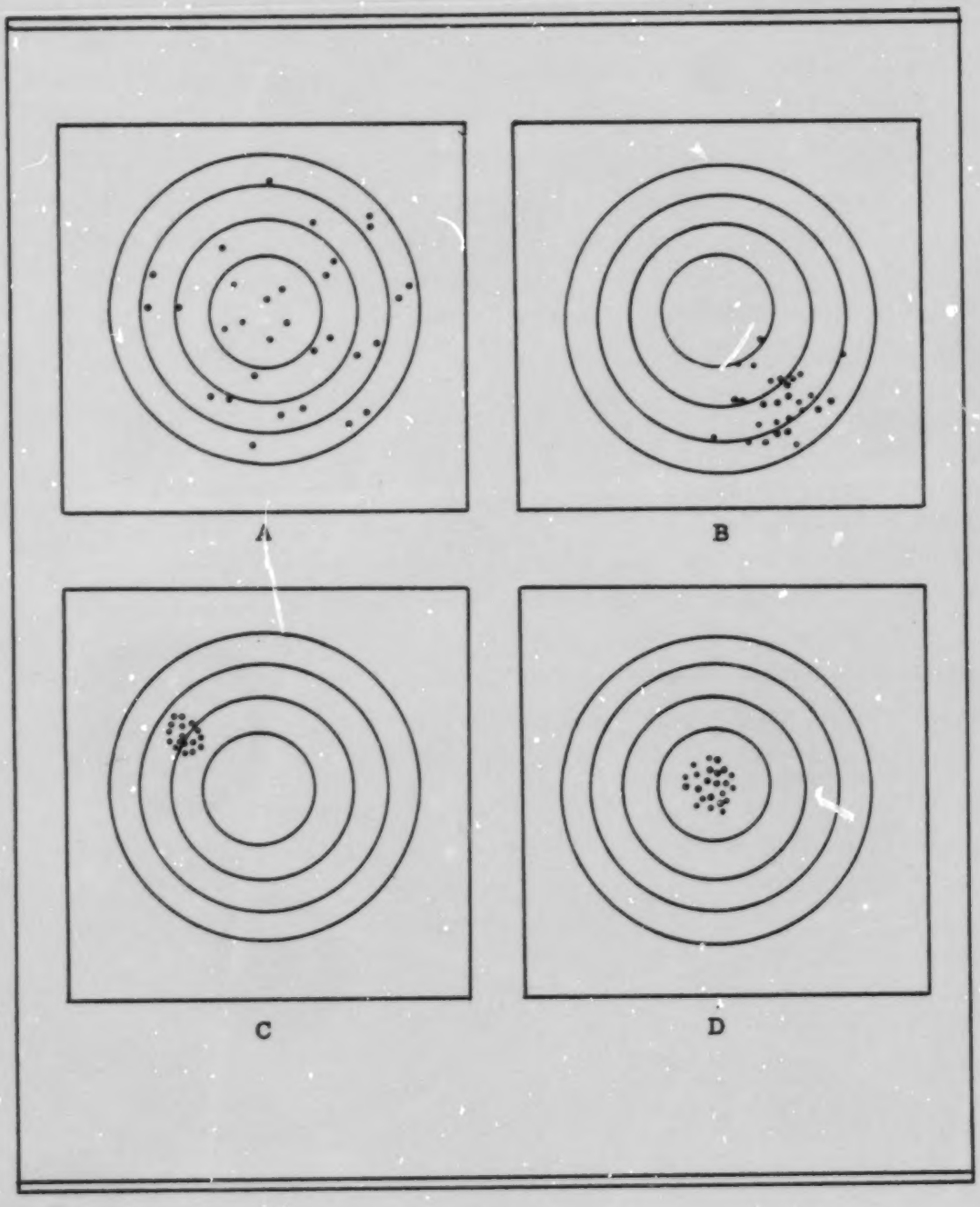
files in a university, it is possible to estimate the percent of procedural discrepancies in the total university. It is not possible to determine the exact number of files containing procedural discrepancies in the university. This means that sample-based estimates of population characteristics are always subject to error. Sampling error is judged in two ways--bias and reliability or precision.

A biased sampling scheme is one which, on repeated trials, produces average estimates of a population characteristic which differs from the true value. An instance of a biased sample would be selection of cases that an auditor has advance reason to believe have overawards. Projecting the results of such a sample to the total population of grant recipients would tend to systematically overestimate the average dollar amount of overawards.

The second major type of sampling error is due to limitations in sample reliability or precision. Very small samples are particularly prone to low reliability. For example, a random sample of 5 student files used to estimate the total amount of student loan overawards in a large university would not be biased because it would not systematically under- or overestimate the true amount of overawards. Such a sample would, however, have a very low reliability in that additional samples of five student files are likely to produce very different population estimates.

The difference between bias and reliability can be illustrated by considering the performance of four guns, A, B, C, and D. Each gun has been fired at a target twenty times. The resulting patterns of hits and misses is shown in Figure 2.1.

FIGURE 2.1: SAMPLING TARGET PRACTICE



Gun A has very low reliability in that its shots are spread all over the target. However, gun A is not biased in that its aim is not systematically low or high or off to the right or left. Gun B has both low reliability and high bias because there is a wide spread of its shots and, on the average, its aim is low and to the right. Gun C has high reliability; its shots all fall in a very limited area. However, its aim is biased high and to the left. Only gun D has both high reliability and no bias as shown by the fact that all the shots are very close to the center of the target.

POINT AND INTERVAL ESTIMATES

When sample data is used to produce a single estimate of a population parameter the estimate is known as a point estimate. Sample based, single value estimates of the number of procedural discrepancies in student aid files, the average level of overawards, and the number of delinquent loans are all examples of point estimates. Because sample-based estimates are inherently subject to a certain error, point estimates seldom exactly match the parameter which they are used to estimate. Therefore, it is a common practice to make interval estimates as well as a point estimate of a parameter.

An interval estimate is one which specifies a range of values rather than a single value. To say that the number of procedural errors falls between 120 and 150 or that the rate of delinquent loans is 4 percent plus or minus 2 percent, or that the average amount of BEOG is $500 plus or minus $50, is to make an interval estimate.

CONFIDENCE INTERVALS, CONFIDENCE LIMITS, AND CONFIDENCE LEVELS

An interval estimate of a population parameter is called a confidence interval and the end points of the interval are known as confidence limits. To understand how statisticians use these terms, it is necessary to define an additional term--confidence level. Confidence level is the level of probability associated with an interval estimate; it is an indicator of the degree of certainty that the particular method of

estimating the confidence interval will produce an estimate which includes the true population value. The higher the confidence level associated with an interval estimate, the more certainty there is that the method of estimation will produce an estimate containing the true value. As an example, if we were to say "On the basis of a random sample of 400 cases, with a 95 percent confidence level, the number of delinquent loans for a lender falls between 45 and 55 percent," what we would be arguing is that if repeated samples of 400 cases each were drawn from the same population, 95 percent of the estimated confidence intervals would contain the true percent of delinquent student loans. The point to remember is that the confidence level refers to the procedure used in drawing the sample and in estimating the confidence interval rather than to any particular interval. Therefore, it is an error to make such statements as "The probability is 95 percent that the percent of delinquent loans is between 45 and 55 percent."

SUMMARY

This completes our introduction to the language of sampling. Statistical sampling, of course, includes many more terms than have been reviewed in this chapter. However, the basic terms and concepts reviewed are sufficient for understanding the advantages and disadvantages of the various sampling designs reviewed in Chapter 3; and for use of the basic sampling formulas introduced in Chapter 4.

3

Types of Samples

There are many types of samples. No single type is superior in all circumstances. This chapter discusses the major options the researcher, auditor and program reviewer have in constructing a sample. The major types of samples are defined, one or more examples of each are given and advantages and disadvantages of each type are discussed. For two of the most common types of samples, simple random and systematic, sections are included on how to construct the sample. The types of samples presented are not mutually exclusive; they can be combined in various ways. It is possible, for example, to draw a multi-stage, stratified, cluster, sequential, dollar-unit, discovery sample.

Simple Random Sampling

Definition: If a sample is drawn from a population in such a way that every possible sample containing the same number of cases has the same chance of being selected, the sampling procedure is called simple random sampling. The most common way of drawing a simple random sample is to assign all cases in a population a number and then select cases by the use of a random number table.

Example: After a program review by OSFA, a large university was required to conduct a sample-based audit of NDSL, SEOG and CWS awards. It was determined that to give a confidence level of 95 percent and a reliability of ± 2 percent, assuming a rate of error in the records of not over 2 percent, a minimum sample size of 137 per program was necessary. To obtain at least 137 students in NDSL, SEOG and CWS, a total sample of 300 student aid recipients was drawn from the university's financial aid computer file using a random number table. In the resulting sample there were 163 CWS recipients, 169 BEOG recipients and 160 NDSL recipients. In this case, a single sample was able to serve the multiple purposes of reviewing awards in three programs.

Advantages: Simple random sampling produces unbiased estimates of population parameters and the results are the easiest to analyze of all statistical sampling methods.

Disadvantages: Simple random sampling requires a complete listing of all cases in the population sampled and, in general, is less precise, given a fixed sample size, than stratified sampling.

Use of a random table number to draw a sample random sample: The most common method of drawing a random sample is through use of a random number table. Tables of random numbers are created in such a way that the integers of 0 through 9 all have an equal probability of occurring in any position on the table. The digits appear on a page in a random fashion. Table 3.1, which follows, is an example of a random number table.

To illustrate use of this table, we will draw a simple random sample of 50 students from a population of 735 student aid recipients at a particular university. First, we take the list of student aid recipients and number them from 001 to 735. Second, we select a starting point on the table. To do so, I simply closed my eyes and stabbed the table with my pencil. The first try missed the table altogether. The second try landed on the '1' underlined on the table. Since we need three-digit numbers (001 to 735), we will consider the '1' to be the first digit and the '2' and '0' digits directly following it to be the second and third digits. The first number is therefore '120'. So student number '120' is selected for inclusion in the sample. We then read down the column to find the next sample member (366). Reading down the column we select '519', '147', and '321'. However, then we come to '827'. Because 827 is not a number on our student list we skip it and continue down the column until we come to the next three digit number between 001 and 735 inclusive.

We continue down the three-digit column selecting eligible numbers, then shift to the next columns of digits reading as far as necessary to draw a sample of 50 students. Table 3.2 includes the actual list of 50 eligible random three-digit numbers selected.

TABLE 3.1: RANDOM NUMBERS

61	81	17	50	68	00	35	10	30	90	59	71	09	95	01	14
78	95	64	65	24	82	14	05	27	63	33	96	10	41	88	70
84	28	44	68	07	47	21	47	56	81	32	87	28	40	40	50
92	33	63	98	99	22	09	21	97	18	10	03	79	46	17	13
15	79	75	50	29	36	12	37	63	39	02	47	57	02	97	17
80	16	09	75	22	28	35	25	53	57	72	64	09	98	63	50
68	20	33	03	43	73	80	96	21	13	97	61	90	37	35	77
55	26	85	04	30	60	68	10	73	53	89	35	58	45	83	23
60	00	37	51	42	89	52	32	46	00	57	02	71	97	44	16
59	69	31	20	16	37	66	34	99	76	07	23	40	85	64	91
84	42	33	66	58	54	17	16	45	73	67	20	09	27	90	96
57	46	65	19	78	34	57	12	77	45	54	65	17	17	30	90
78	17	51	47	69	22	41	48	01	99	66	46	00	28	21	74
27	66	33	21	49	11	24	15	33	70	06	95	04	67	98	56
82	54	98	27	81	86	77	35	87	56	32	72	60	90	26	75
33	06	79	71	73	57	96	74	85	94	36	97	87	79	82	00
77	94	61	11	69	61	78	78	36	51	45	21	82	94	39	22
87	15	49	66	56	55	34	99	05	26	45	35	59	83	55	47
24	98	52	45	79	85	15	67	32	21	29	94	98	90	02	27
05	66	15	23	83	66	24	98	06	75	60	69	64	26	58	24
84	90	70	29	01	36	90	78	56	40	61	00	58	40	75	37
49	50	30	71	87	38	70	10	80	71	12	54	60	76	62	13
27	53	95	47	04	78	61	85	56	15	71	76	25	31	96	39
56	17	07	83	96	29	88	39	67	86	98	23	95	03	82	62
41	67	05	42	29	18	54	76	71	82	04	81	82	63	00	23

TABLE 3.2: SELECTED SAMPLE CASES

120	547	163	035	124
366	542	585	214	677
519	680	692	721	178
147	248	491	209	534
327	074	735	612	515
111	293	696	380	624
245	222	565	068	690
523	437	013	417	510
029	306	047	457	405
071	428	291	241	147

When we examine the list of 50 numbers, we see that the number 147 has been drawn twice. For a simple random sample, the technically correct procedure in such cases is to count the data collected from student number 147 twice in the analysis. This procedure is called sampling with replacement. In actual practice, most researchers simply draw additional cases until they reach the desired sample size and count each case once in the analysis. This procedure is called sampling without replacement.

Stratified Sampling

Definition: A stratified sample is one obtained by separating the population into nonoverlapping groups called strata and then selecting a simple random sample from each stratum.

Example: For a review of NDSL loans made by a major lender, the reviewer divided the population of loans into two primary strata; loans to students currently in school and loans to students who are no longer in school. The second stratum was further subdivided into loans in the grace period, loans that have been repaid, delinquent loans, and loans in the process of repayment. From each of the resulting five strata, a simple random sample was drawn, and the selected cases were reviewed. This procedure guaranteed the reviewers an adequate number of loans within each stratum to make objective statements about members of the stratum and make projections to the total population of loans.

Advantages: Stratified sampling produces unbiased population estimates and it is more precise than simple random sampling given a fixed sample size. Stratified sampling introduces a great deal of flexibility into statistical sampling. Members of groups of special interest can be given a higher probability of being sampled than members of groups of low interest. The selector's prior knowledge can be incorporated in the sample design.

Disadvantages: Stratified sampling requires advance knowledge of the proportion of the population in each stratum; otherwise, the precision of the sample is decreased. If lists of cases in each stratum are not available, stratified sampling may not be possible.

Cluster Sampling

Definition: A cluster sample is a simple random sample in which the sampling units are collections or clusters of cases.

Example: For a review of the CWS program in a university, a team of program reviewers sampled work-study time sheets by selecting three weeks at random and then reviewing all the time sheets for each of the three weeks selected. Because the time sheets were selected in groups rather than individually this sampling method is called cluster sampling. Cluster sampling is commonly used when the population to be sampled is dispersed over a wide geographical area. For example, to reduce travel costs for a follow-up study of student loan defaults, several cities and towns were selected at random and then loan recipients sampled within the selected areas.

Advantages: Cluster samples produce unbiased estimates of population parameters. They require a listing of only those cases included in the selected clusters. In many circumstances cluster sampling is more cost-effective than other methods. For personal interview surveys conducted over a wide area, travel costs can often be substantially reduced by clustering. When complete population lists are not available, clustering can reduce costs in sample selection.

Disadvantages: Clustering generally produces less precise population estimates given a fixed sample size than simple random or stratified sampling. It is usually not possible to determine in advance the minimum sample size required to achieve a predetermined level of precision. Computation of population estimates from cluster samples can be quite complex.

Systematic (Interval) Sampling

Definition: Systematic sampling is a method of selection whereby sample cases are drawn from a population at some fixed interval but where the starting case is selected at random.

Example: For an audit of BEOG awards an auditor determined that a minimum sample of 100 awards was necessary. The university being audited had a validation roster containing 1,172. The audit reviewer divided 1,172 by 100 to obtain 11.72, which he rounded down to the number 11. Selecting the random number of 3 as a starting point, he selected for review the 3rd name on the validation roster, the 14th name (3 + 11), the 25th name (14 + 11) and so on until he had worked through the entire list.

Advantages: Often systematic samples are the easiest type to construct. They do not require a complete listing of all cases. In most circumstances, when the cases are ordered at random, or in alphabetical order, the resulting sample is unbiased. Certain methods of ordering cases such as date of loan or size of loan, can introduce implicit stratification into the sample and thereby increase precision.

Disadvantages: Systematic sampling is not usable when cases are missing from the files or records. Periodic ordering of files or records can introduce bias into systematic sampling. For example, if a particular record system added a new file for every work day, any systematic sample of the files which had '5' as a factor of its sampling interval would result in selection of cases all from the same day of the week.

Drawing a systematic sample from filing cabinets: If the program under review can provide access to filing cabinets containing the files of all student aid recipients, systematic sampling from the filing cabinets may be appropriate.

Procedures for sampling from filing cabinets	Line Number
1. Determine the total number of student aid recipients	N = _____(1)
2. Using form C on page 41 determine the minimum necessary sample size	n = _____(2)
3. Divide line 1 by line 2	N/n = _____(3)
4. Truncate line 3 to an integer (For example, if line 3 equals '17.7' write '17' on line 4) Sampling Interval	= _____(4)
5. Employing Table 3.1 on page 15 select a random number between 0 and 9 Random Number	= _____(5)

6. Using Table 3.3 below select a starting number.

7. Start with the top drawer of the first filing cabinet. Count file folders until you get to the starting number on line 6. Select this case for review.

8. Starting with the last file selected, count forward the number of files specified by the sampling interval (line 5). Select the file obtained for review.

9. Repeat step 8 until you have worked your way through all the files.

TABLE 3.3: RANDOM START FOR INTERVAL SAMPLING

Sampling Interval (Line 4)	Number on Line 5									
	0	1	2	3	4	5	6	7	8	9
2	2	2	2	1	1	2	2	2	1	2
3	2	1	2	3	2	2	1	2	2	3
4	1	2	4	2	1	3	1	1	1	1
5	4	2	4	2	5	4	3	1	4	1
6	6	5	3	1	5	2	2	6	3	2
7	5	5	7	5	4	4	2	2	1	5
8	5	6	8	5	2	2	5	2	8	7
9	7	5	6	6	5	3	9	3	8	9
10 - 11	6	5	1	3	5	3	7	5	6	2
12 - 13	7	1	6	11	12	4	8	8	4	9
14 - 16	9	1	8	8	3	10	2	5	6	12
17 - 19	15	11	17	11	6	17	6	5	10	3
20 - 24	7	10	17	14	9	5	7	4	5	8
25 - 29	1	23	25	5	5	11	4	24	10	6
30 - 35	30	9	8	6	17	19	21	29	17	25
36 - 41	32	36	15	28	13	3	22	17	19	33
42 - 49	25	37	12	30	30	38	26	31	9	11
50 - 57	18	16	30	9	15	31	8	24	17	23
58 - 69	4	29	30	14	3	6	35	31	25	38
70 - 81	38	1	14	19	11	52	44	45	53	10
82 - 99	78	28	17	76	74	60	37	34	8	13
100+	79	84	18	43	72	24	15	33	59	11

Starting Number = _____(6)

Dollar-Unit Sampling

Definition: When the sampling unit is defined as an individual dollar rather than an individual loan, grant, etc., the sample is called a dollar-unit sample.

Example: Sampling loans or accounts tends to give equal weight to each loan or account. For many purposes, a one hundred dollar loan should not be counted equally with a ten thousand dollar loan. This is particularly true when the goal of the auditor or reviewer is to estimate dollar amounts of overpayments or discrepancies. One solution to this problem is dollar unit sampling. Rather than treating loans, grants or recipients as sampling cases, dollar unit sampling uses the dollars involved as the sampling unit. As an illustration, consider an audit review of a university which administered 483 CWS awards totalling $724,500. The university supplied the auditor with a list of award recipients and the dollar amount of each loan. Because the awards varied greatly in amount, the auditor decided to conduct a systematic, dollar-unit sample. Having determined that a sample size of 142 was necessary to her purposes, she divided 724,500 by 142 to arrive at a sampling interval of 5102. She then selected a random starting point of 1847 in the first interval.

Working her way through the list of loans, she made a running table of the amount of loans on the listing, selecting for review the loan containing the 1847th dollar, the 6949th dollar, (1847 + 5102) the 12051th dollar (6949 + 5102) and so on. Using this method of sampling, each loan had a probability of being selected directly proportional to its size. Thus a loan of $1000 had twice the chance of being selected as a loan of $500.

Advantages: Dollar-unit sampling produces unbiased population estimates. It produces more precise estimates of dollar amount population parameters than loan, grant, or recipient sampling given a fixed sample size. It is an effective way of locating large errors clustered in large accounts that are almost impossible to detect by account sampling. Finally, the problems of converting error frequencies into dollar amounts for population projections are avoided.

Disadvantages: Dollar-unit sampling produces less precise estimates of error frequencies in a population than loan, grant, or recipient sampling given a fixed sample size. In many cases, the data on account size required to draw a dollar-unit sample are not available in advance.

Sequential (Stop or Go) Sampling

Definition: In sequential sampling on the basis of a minimum initial sample, decisions are made as to whether it is necessary to sample additional cases, and if so, what type of cases should be sampled.

Example: During a program review, a team of reviewers drew a minimum initial sample of 50 student files from a population of 14,187 student aid recipients. From a review of the initial sample they found no errors in SEOG and NDSL awards and grants. However, they discovered a substantial number of discrepancies in CWS awards. On the basis of this information, they decided to terminate their audit of SEOG and NDSL and to draw an additional sample of 50 CWS awards.

Advantages: Sequential sampling allows minimization of sample size, does not require advance knowledge of population distributions, allows modification of sample design to take advantage of knowledge gained during the previous sequence and adds flexibility to discovery sampling designs.

Disadvantages: Sequential sampling may be very time consuming because it requires that the sampling process be periodically halted to analyze data gathered.

Discovery (Exploratory) Sampling

Definition: Discovery sampling is a sampling design used to locate examples or establish a maximum rate for infrequent occurrences. Discovery sampling is a method of giving assurance to an auditor or program reviewer that if some critical event has occurred with some minimum frequency, the sample will contain at least one example of this event.

Example: Suppose an auditor wished to examine 20,000 grant vouchers for possible cases of fraud. To assure that there were no cases of fraud he/she would, of course, be required to examine all 20,000 vouchers, which might not be practical. One alternative to a 100 percent review would be to sample enough cases to assure that if fraud did exist at above a certain level or rate the auditor will have reasonable certainty of discovering at least one case. If the auditor drew a random sample of 300 vouchers, he could be assured at a 95 percent confidence level, that if fraud occurred in 1 percent or more of the loans at least one case of fraud would be included in the sample. Therefore, if an examination of the sample vouchers revealed no examples of fraud, the auditor could reasonably conclude that even if fraud did occur, it occurred in less than 1 percent of the loans. Discovery sampling is often used with sequential sampling. After review of the initial sample a decision is made concerning the need for examination of additional cases.

Advantages: Discovery sampling is an effective method of establishing maximum rates for rare but significant events.

Disadvantages: Discovery sampling provides no basis for making population estimates once a discrepancy is discovered; therefore, it is most useful in conjunction with sequential sampling. After an initial minimum sample is drawn, sampling is stopped if no error is discovered. The auditor or reviewer continues the sampling if a discrepancy is discovered.

Multi-Stage Sampling

Definition: Multi-Stage sampling is a process of selecting a sample in two or more successive and contingent stages.

Examples: There is no complete listing of college students in the United States and therefore a simple random sample of college students is not possible. One way to draw a representative sample would be to first sample colleges which do have complete student rosters, and then sample students attending the selected colleges. The precision of the sampling design could be improved by first stratifying colleges by such variables as size, type, and geographic location and then sampling from each stratum. Within colleges, the student population could also be stratified by year-in-school, enrollment status, sex, race, and so on.

At times, samples can involve many stages. A recent survey of elementary school children first sampled school districts, then elementary schools within the selected districts, then classes within the selected schools, then students within the selected classrooms. At every stage, the sample was stratified and weighted to improve precision.

Advantages: Multi-stage sampling introduces a great deal of flexibility into sample design and makes possible sampling of populations for which there are no complete lists of cases. It also can incorporate the advantages of stratification and clustering in a single sample.

Disadvantages: Multi-stage sampling introduces great complexities into data analysis. Estimation of confidence intervals for a multi-stage sample usually requires knowledge of advanced statistical procedures.

Opportunity Sampling

Definition: Opportunity sampling is the selection of sample cases in a haphazard way. The selector takes an opportunity sample when he selects any case he happens to run across for inclusion in the sample.

Examples: Samples of the first twenty files in a cabinet of student financial aid recipients for review is an example of opportunity sampling. A common version of opportunity sampling is "man-on-the-street" interviews conducted by television, radio and newspapers as an informal measure of public opinion on current events.

Advantages: Opportunity sampling is an easy sample selection method because it imposes no constraints on which cases may be selected.

Disadvantages: Opportunity sampling potentially introduces bias into the sample because there is no way of being certain that the sampled cases are truly reflective of the total population sampled.

For example, selection of the first twenty student files in a cabinet may result in a sample limited to recent aid recipients. "Man-in-the-street" interviews conducted during working hours may exclude working people from the sample. Therefore, the results of opportunity sampling cannot provide the basis for objective projection of sample results to the total population.

Quota Sampling

Definition: When a pre-specified number or quota of sample cases are selected on an opportunity basis from the various groups or categories which compose the population under study, the sample is called a quota sample. A quota sample is a stratified opportunity sample.

Example: If a researcher wished to sample the student population of a university in which 30 percent were freshmen; sophomores, juniors and seniors each composed 20 percent of the students; and 10 percent were graduate students, he might select the first 30 freshmen, 20 sophomores, 20 juniors, 20 seniors and 10 graduate students leaving the student union.

Advantages: Quota sampling increases the representativeness of opportunity sampling and thereby potentially reduces bias.

Disadvantages: Quota sampling rests on the false assumption that membership in a category automatically qualifies a case to represent all members of that category. Quota sampling is subject to selection bias and has unknown statistical properties. Therefore, the results of quota sampling cannot provide the basis for objective projection of sample results to the total population.

SUMMARY

Choosing the Right Sample Design

There are no simple rules for choosing the optimal sample design that apply in all circumstances. However, several basic guidelines can be used. Simple random or systematic sampling are most effective when complete lists of the population exist and the researcher desires a sampling design that lends itself to simple analysis. Stratified sampling is advantageous when the researcher wishes to assure inclusion in the sample of certain subpopulations or to increase the efficiency of simple random sampling. Cluster sampling can save costs for personal interview survey conducted over a wide area, or when complete population lists are not available. Dollar unit sampling is advantageous when auditing a system of records containing many small and a few large accounts. Sequential and discovery sampling are most useful when investigating a population about which there is little advance knowledge.

Ultimately study goals and resources dictate choice of sample design. The great variety of sample design choices permit tayloring of a sample to many different study purposes and budgets.

4

Computing Sample Statistics

This Chapter contains a series of forms to aid in calculating a variety of common sample statistics. Form A is for calculating the estimate of the population variance from a sample. Form B can be used in developing population estimates from a simple random sample and Form C is for determining minimum necessary sample sizes. Each of the forms contains a step-by-step procedure for calculating these important sample statistics.

In addition to these hand calculation forms, two forms have been provided for computing these statistics with a calculator. Form D shows the procedure for using a calculator to determine population variance, and Form E, for developing population estimates. It should be noted that while a calculator can be used for all the sample statistics described in this manual, efficiencies can be gained in its use for calculating population variance and population estimates.

FORM A

Estimating the Variance of a Population from a Sample (s^2)

$$\frac{\Sigma (X_i - \bar{X})^2}{(n-1)} = \frac{\Sigma X_i^2 - (\Sigma X_i)^2/n}{(n-1)}$$

STEP	LINE NUMBER
A. How many cases are in the sample?	$n =$ _____ (1)
B. Subtract 1 from line 1	$n - 1 =$ _____ (2)

C. Is the variable a categorical variable (such as sex or recipient/nonrecipient) or a continuous variable (such as income or age)?

- ☐ Categorical — Go to Step G
- ☐ Continuous — Continue with Step D

D. Calculate the numerator: the "sum of squared deviations"

$$\Sigma(X_i - \bar{X})^2 = \Sigma X_i^2 - (\Sigma X_i)^2/n$$

D1 Square the value of each sample case and sum the results

$$X_1^2 + X_2^2 + X_3^2 \cdots X_n^2 = \Sigma X_i^2 =$$ _____ (3)

D2 Add the values from all the cases in the sample together

$$X_1 + X_2 + X_3 \cdots X_n = \Sigma X_i =$$ _____ (4)

D3 Square line 4

$$(\Sigma X_i)^2 = (\Sigma X_i) \cdot (\Sigma X_i) =$$ _____ (5)

D4 Divide line 4 by line 1

$(\Sigma X_i)^2/n =$ _____(6)

D5 Subtract line 6 from line 3

$\Sigma X_i^2 - (\Sigma X_i)^2/n =$ _____(7)

E. Calculate the estimate of the population variance. Divide line 7 by line 2

$$\frac{\Sigma X_i^2 - (\Sigma X_i)^2/n}{(n-1)} =$$ _____(8)

F. How many cases in the sample are in the category of interest?

Number of Cases in Category = f = _____(9)

(For example, if you are interested in estimating the variance of sex, how many females are there in the sample? For docotomous variables it makes no difference which category is chosen.)

G. Square line 9

$f^2 = f \cdot f =$ _____(10)

H. Divide line 10 by line 1

$f^2/n =$ _____(11)

I. Subtract line 11 from line 9

$f - f^2/n =$ _____(12)

J. Divide line 12 by line 2

$$\frac{f - f^2/n}{(n-1)} = \frac{\Sigma(X_i - \bar{X})^2}{(n-1)} =$$ _____(13)

FORM B

Developing Population Estimates From a Simple Random Sample

STEP			LINE NUMBER
A.	**Basic Sample Information**		
A1	How many cases are in the sample?	$n =$	_____ (1)
A2	How many cases are in the total population from which the sample was drawn?	$N =$	_____ (2)
B.	**Calculate the sample mean ($\bar{x}$)**		
B1	Add the values for all the cases in the sample together	$x_1 + x_2 + x_3 \ldots + x_n = \Sigma x_i =$	_____ (3)
B2	Divide line 3 by line 1	$\Sigma x_i / n = \bar{x} =$	_____ (4)
C.	**Calculate the estimate of the sampling mean standard deviation** $\hat{\sigma}_{\bar{x}} = \sqrt{\frac{s^2}{n} \cdot \frac{N-n}{N}}$		
C1	Using form A, calculate the estimated sample variance	$s^2 =$	_____ (5)
C2	Divide line 5 by line 1	$s^2/n =$	_____ (6)
C3	Subtract line 1 from line 2	$N - n =$	_____ (7)
C4	Divide line 7 by line 2	$\frac{N-n}{N} =$	_____ (8)
C5	Multiply line 6 by line 8	$\frac{s^2}{n} \cdot \frac{N-n}{N} =$	_____ (9)

C6 Take the square root of line 9 $\sqrt{\frac{s^2}{n} \cdot \frac{N-n}{N}} =$ _____ (10)

D. Set the "confidence level" (confidence level is defined on page 13)

Confidence level = CL = _____(11)

E. Determine the "Z" or "K" value from the table below

If CL =	If $n \geq 30$ then Z =	If $n < 30$ then K =
80%	1.28	2.24
90%	1.64	3.16
95%	1.96	4.44
99%	2.55	10.00

K or Z = _____(12)

F. Calculate the estimate of the population total

$\hat{\tau} = N \cdot \bar{x}$

Multiply line 2 by line 4 $\hat{\tau} = N \cdot \bar{x} =$ _____(13)

G. Calculate the confidence interval of the estimated population total

Multiply line 2 by line 12 by line 10

If $n \geq 30$ $CI = N \cdot Z \cdot \hat{\sigma}\bar{x} =$ _____(14)

OR

If $n < 30$ $CI = N \cdot K \cdot \hat{\sigma}\bar{x} =$ _____(14)

H. Calculate the upper bound of the confidence interval

$\hat{\tau} + CI$

Add line 14 and 13 $\hat{\tau} + CI =$ _____(15)

I. Calculate the lower bound of the confidence interval

$$\hat{\tau} - CI$$

Subtract line 14 from line 13 $\hat{\tau} - CI =$ _____(16)

J. Interpreting the results

Fill in the blanks in the sentence below.

"On the basis of a sample of _____ (line 1) cases, it can be estimated with _____ (line 11) percent confidence, the total value of _____ (variable name) for the population sampled falls between _____ (line 16) and _____ (line 15) with the most likely value _____ (line 13).

FORM C

Determining Sample Sizes (n)

STEP	LINE NUMBER

A. Is the variable to be estimated a categorical variable (such as sex or percentage of errors) **or a continuous variable** (such as total dollars expended or average cost)?

- ☐ Categorical variable - Go to Step H
- ☐ Continuous variable - Go to Step B

B. Establish the *average* acceptable error.

E = _____(1)

(If, for example, you wish to estimate the *average* weight of students in a class of thirty within two pounds write "2" on line 1. However, if you wish to estimate the *total* weight of students in the class within twelve pounds, you must first calculate the average acceptable error (E) by dividing the total acceptable error (TE) by the number of cases in the population to be sampled. In this case the average error would be:

$$E = \frac{TE}{N} = \frac{12}{30} = .4$$

Write the result on line 1)

C. Set the "confidence level"

Confidence level = _____(2)

(See page 13 for a definition of confidence level)

D. Determine the "Z" value from the table below

If CL =	than Z =
80%	1.28
90%	1.64
95%	1.96
99%	2.58

Z = _____(3)

E. How many cases are there in the total population from which the sample is to be drawn?

N = _____(4)

F. Determine the estimated population variance ($\hat{\sigma}^2$) (see page A4 for a definition of variance)

$\hat{\sigma}^2$ can be estimated from:

1. Past experience
2. Pilot study
3. If sampling from an approximately normally distributed population, the variance can roughly be estimated as:

$$\hat{\sigma}^2 = \frac{R^2}{25}$$

where R is the range. The range is the highest value minus the lowest value.

$\hat{\sigma}^2$ = _____(5)

G. Calculate the minimum necessary sample size.

$$n = \frac{NZ^2 \cdot \hat{\sigma}^2}{E^2N + Z^2\hat{\sigma}^2}$$

G1 Square line 3

Multiply line 3 by itself. $Z^2 =$ _____(6)

G2 Multiply line 5 by line 6

$\hat{\sigma}^2 Z^2 =$ _____(7)

G3 Multiply line 7 by line 4

$\hat{\sigma}^2 Z^2 N =$ _____(8)

G4 Square line 1

$E^2 =$ _____(9)

G5 Multiply line 9 by line 4

$E^2 N =$ _____(10)

G6 Add lines 10 and 7

$E^2 N + \overline{\hat{\sigma}^2 Z^2} =$ _____(11)

G7 Divide line 8 by line 11

$$n = \frac{\hat{\sigma}^2 \cdot NZ^2}{E^2 N + \hat{\sigma}^2 Z^2} =$$ _____(12)

If line 12 is less than 30; sample a minimum of 30 cases

H. If the variable to be estimated is a categorical variable:

Establish the *proportion* acceptable error

$E =$ _____(13)

(For example, if you wish to estimate the *proportion* of students in a class of thirty who are female with .05 or less error write .05 on line 13. However, if you wish to estimate the *total number* of students who are female within three students, you must first calculate

the proportion acceptable error, E, by dividing the total acceptable by the number of cases in the population to be samled. In this case the proportion error would be

$$E = \frac{TE}{N} = \frac{3}{30} = .10$$

Write the result on line 13)

I. Set the "confidence level"

Confidence level = CL = _____(14)

(See page 13 for a definition of confidence level)

J. Determine the "Z" value from the table below

If CL =	then Z =
80%	1.28
90%	1.64
95%	1.96
99%	2.58

Z = _____(15)

K. Determine the estimated population percentage for the category to be estimated. (P)

P can be estimated from:

1. Past experience
2. A pilot study
3. Assuming the "maximum variance" and setting P = .5

P = _____(16)

L. How many cases are there in the total population from which the sample is to be drawn?

N = _____(17)

M. Calculate the minimum necessary sample size

$$n = \frac{NZ^2 \cdot P(1-P)}{E^2N + Z^2P(1-P)}$$

M1 Square line 15 (Z^2)
(Multiply line 15 by itself)

$Z \cdot Z = Z^2 =$ _____ (18)

M2 Subtract line 16 from 1

$1 - P =$ _____ (19)

M3 Multiply line 16 by line 19

$P(1-P) =$ _____ (20)

M4 Multiply line 18 by line 20

$Z^2 \cdot P(1-P) =$ _____ (21)

M5 Multiply line 21 by line 17

$NZ^2 \cdot P(1-P) =$ _____ (22)

M6 Square line 13

$E \cdot E = E^2 =$ _____ (23)

M7 Multiply line 23 by line 17

$E^2 \cdot N =$ _____ (24)

M8 Add lines 21 and 24

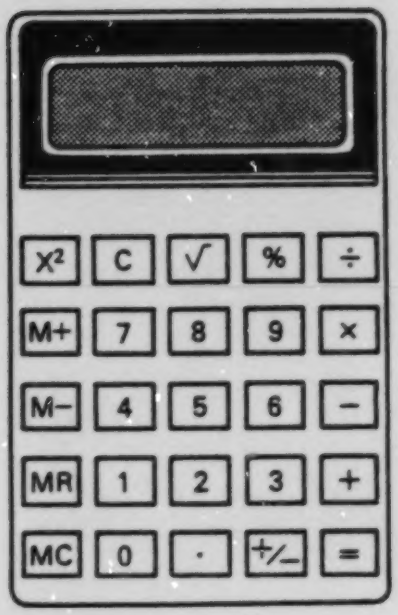
$E^2N + Z^2 \cdot P(1-P) =$ _____ (25)

M9 Divide line 22 by line 25

$$n = \frac{NZ^2 \cdot P(1-P)}{E^2N + Z^2 \cdot P(1-P)} =$$ _____ (26)

If line 26 is less than 30, sample a minimum of 30 cases

USING A CALCULATOR TO COMPUTE STANDARD SAMPLING STATISTICS



Forms D and E present a method of using a hand calculator to compute population estimates from a sample. To use the forms requires a hand calculator with memory, square, and square-root keys and which employs "standard algebraic hierarchy;" i.e., squares and square-roots are performed as soon as the appropriate keys are pressed, and multiplication and division are performed before addition and subtraction. To test if your calculator conforms to standard algebraic hierarchy, press the following keys:

[3] [+] [4] [√] [×] [5] [=]

If the display shows '13' the following forms are usable with your calculator.

FORM D

Using a Calculator to Compute the Estimate of the Population Variance from a Sample

Basic Information

How many cases are there in the sample? $n =$ _____(1)

Computation

Enter	Press			Display	Comments
	C	MC		0	Clear memory and display
X_1	M+	X^2	+	X_1^2	Sum values of X_i^2 in display and sum values of X_i in memory
X_2	M+	X^2	+	$X_1^2 + X_2^2$	
X_i	M+	X^2	+	$X_1^2 + X_2^2 + X_3^2$	Continue for all cases
X_n	M+	−		ΣX_i^2	Sum of the squared value of all cases
	MR			ΣX_i	Sum of the values of all cases Write contents of display on line 2 $\Sigma X_i =$ _____(2)
	X^2	+		$(\Sigma X_i)^2$	
n (From Line 1)	MC	M+	=	$\Sigma(X_i - \bar{X})^2$	Sum of the squared deviations Write contents of display on line 3 $\Sigma(X_i - \bar{X})^2 =$ _____(3)
1	M−				
$\Sigma(X_i - \bar{X})^2$ (From Line 3)	+	MR	=	s^2	The estimated variance of the population sampled

FORM E

Using a Calculator to Compute Population Estimates from a Simple Random Sample

Basic Information

How many cases are there in the sample? $n =$ _____ (1)

How many cases are there in the population from which the sample was drawn? $N =$ _____ (2)

What is the estimated variance of the population sampled?
(Use form D to calculate the estimated population variance) $s^2 =$ _____ (3)

What is the sum of the values for all cases in the sample (Line 2 from form D) $\Sigma X_i =$ _____ (4)

Set the "confidence level"
(confidence level is defined on page 13)

Confidence level = CL = _____ (5)

Determine the "Z" or "K" value from the table below

If CL =	If $n \geq 30$ then Z =	If $n < 30$ then K =
80%	1.28	2.24
90%	1.64	3.16
95%	1.96	4.44
99%	2.55	10.00

K or Z = _____ (6)

Computation

Enter	Press	Display	Comments
	C MC	0	Clear Memory and Display
N (from line 2)	M+ −	N	

Enter	Press	Display	Comments
n (from line 1)	= + MR ×	$\frac{N-n}{N}$	Finite population correction
s^2 (from line 3)	+		
n (from line 1)	= √ ×	$\hat{\sigma}_{\bar{X}}$	Sampling mean standard deviation
Z or K (from line 6)	× MR = MC M+	$N \cdot Z \cdot \hat{\sigma}_{\bar{X}}$	
ΣX_i (from line 4)	+	ΣX_i	
n (from line 1)	×	$\bar{X}$	Mean sample value
N (from line 2)	+	$\hat{\tau}$	Estimate of population total Enter contents of display on line $\hat{\tau}$ = _____(7)
	MR =	$\hat{\tau} + N \cdot Z \cdot \hat{\sigma}_{\bar{X}}$	Upper bound of confidence interval Enter contents of display on line 8 $\hat{\tau} + N \cdot Z \cdot \hat{\sigma}_{\bar{X}}$ = _____(8)
	− MR × 2 =		Lower bound of confidence interval Enter contents of display on line 9 $\hat{\tau} - N \cdot Z \cdot \hat{\sigma}_{\bar{X}}$ = _____(9)

Interpreting the Results

Fill in the blanks in the sentence below.

On the basis of a sample of _____ cases, it can be estimated with _____
line 1 *line 5*
percent confidence, the total value of _____ for the population
variable name
sampled falls between _____ and _____ with the most likely value _____.
line 9 *line 8* *line 7*

5

Applications of Sampling to Student Financial Aid

This chapter contains three examples of the potential uses of sampling statistics by the Office of Student Financial Assistance. Although the details of the examples are fictionalized, they are all based on a combination of actual cases. The examples are designed to both illustrate the application of statistical sampling in OSFA and to address a range of potential problems that could arise in those applications.

EXAMPLE 1: REVIEW OF SEOG AWARDS AT UNIVERSITY A

A financial aid program review at University A revealed five overawards in the twenty-five SEOG awards reviewed. As a result, the Department required that the University either conduct a complete audit of their SEOG awards or perform its own statistically sound and representative sample and project the results of the sample to the total SEOG population during the period of the audit. The University selected the latter option and proposed a 10% simple random sample of the 1460 SEOG awards made during the period of the audit. In evaluating the proposed sample the Department determined that the sample would have to be sufficient to estimate the number of overawards within ± 50 and the amount of overawards ± 50,000, at a 95 percent confidence level. To determine whether the University's proposed sample plan met these criteria, Form C from this manual was used. A copy of the completed form with relevant comments is attached.

FORM C

Determining Sample Sizes (n)

STEP — **LINE NUMBER**

A. Is the variable to be estimated a categorical variable (such as sex or percentage of errors) or a continuous variable (such as total dollars expended or average cost)?

☒ Categorical variable - Go to Step H

☐ Continuous variable - Go to Step B

B. Establish the *average* acceptable error.

(If, for example, you wish to estimate the *average* weight of students in a class of thirty within two pounds write "2" on line 1. However, if you wish to estimate the *total* weight of students in the class within twelve pounds, you must first calculate the average acceptable error (E) by dividing the total acceptable error (TE) by the number of cases in the population to be sampled. In this case the average error would be:

$$E = \frac{TE}{N} = \frac{12}{30} = .4$$

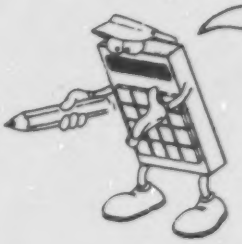
Write the result on line 1)

E = _____(1)

C. Set the "confidence level"

Confidence level = _____(2)

(See page 9 for a definition of confidence level)



In this case, the sample will be used to estimate both the dollar amount of overpayments, (a continuous variable) and the number of overpayments (a categorical variable).

For such multiple use samples, the necessary sample size should be determined independently for each use and the largest resulting estimate of minimum necessary sample size used.

In this example we will first estimate the necessary sample size needed to determine the number of SEOG overawards.

BEST COPY AVAILABLE

G1 Square line 3

Multiply line 3 by itself. $Z^2 =$ _____ (6)

G2 Multiply line 5 by line 6

$\hat{\sigma}^2 Z^2 =$ _____ (7)

G3 Multiply line 7 by line 4

$\hat{\sigma}^2 Z^2 N =$ _____ (8)

G4 Square line 1

$E^2 =$ _____ (9)

G5 Multiply line 9 by line 4

$E^2 N =$ _____ (10)

G6 Add lines 10 and 7

$E^2 N + \hat{\sigma}^2 Z^2 =$ _____ (11)

G7 Divide line 8 by line 11

$$n = \frac{\hat{\sigma}^2 \cdot NZ^2}{E^2 N + \hat{\sigma}^2 Z^2} =$$ _____ (12)

If line 12 is less than 30; sample a minimum of 30 cases

H. **If the variable to be estimated is a categorical variable:**

Establish the *proportion* acceptable error

$E =$ .034 (13)

The level of acceptable error was decided to be ± 50 overawards. To convert this figure into proportion acceptable error, the following formula was used: $E = \frac{TE}{N} = \frac{50}{1460.} = .034$

(For example, if you wish to estimate the *proportion of* students in a class of thirty who are female with .05 or less error write .05 on line 13. However, if you wish to estimate the *total number* of students who are female within three students, you must first calculate

BEST COPY AVAILABLE

the proportion acceptable error, E, by dividing the total acceptable by the number of cases in the population to be samled. In this case the proportion error would be

$$E = \frac{TE}{N} = \frac{3}{30} = .10$$

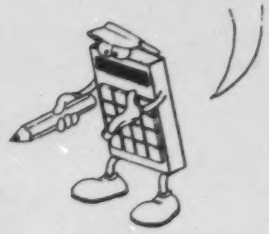
Write the result on line 13)

I. **Set the "confidence level"**

Confidence level = CL = 95% (14)

(See page for a definition of confidence level)

95% is the confidence level selected by the Department.



J. **Determine the "Z" value from the table below**

If CL =	then Z =
80%	1.28
90%	1.64
95%	1.96
99%	2.58

Z = 1.96 (15)

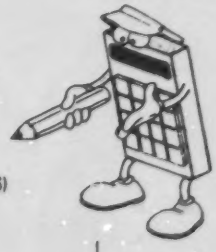
K. **Determine the estimated population percentage for the category to be estimated. (P)**

P can be estimated from:

1. Past experience
2. A pilot study
3. Assuming the "maximum variance" and setting P = .5

P = .20 (16)

The initial review found 5 overawards out of 25 files reviewed or .20 overawards.



L. **How many cases are there in the total population from which the sample is to be drawn?**

N = 1460 (17)

BEST COPY AVAILABLE

M. Calculate the minimum necessary sample size

$$n = \frac{NZ^2 \cdot P(1-P)}{E^2N + Z^2P(1-P)}$$

M1 Square line 15(Z^2)
(Multiply line 15 by itself)

$Z \cdot Z = Z^2 =$ 3.84 (18)

M2 Subtract line 18 from 1

$1 - P =$ .80 (19)

M3 Multiply line 18 by line 19

$P(1-P) =$ .16 (20)

M4 Multiply line 18 by line 20

$Z^2 \cdot P(1-P) =$ .6144 (21)

M5 Multiply line 21 by line 17

$NZ^2 \cdot P(1-P) =$ 897 (22)

M6 Square line 13

$E \cdot E = E^2 =$ .000117 (23)

M7 Multiply line 23 by line 17

$E^2 \cdot N =$ 1.7077 (24)

M8 Add lines 21 and 24

$E^2N + Z^2 \cdot P(1-P) =$ 2.322 (25)

M9 Divide line 22 by line 25

$n = \frac{NZ^2 \cdot P(1-P)}{E^2N + Z^2 \cdot P(1-P)} =$ 386 (26)

If line 26 is less than 30, sample a minimum of 30 cases

BEST COPY AVAILABLE

Therefore, the minimum sample size necessary to estimate the total number of overawards ± 50 is 386. To determine the sample necessary to estimate the total amount of overawards ± $50,000 the equation for minimum sample size in Appendix B was used directly.

$$n = \frac{NZ^2 \cdot \hat{\sigma}^2}{E^2N + Z^2\hat{\sigma}^2}$$

In our example:

N = 1460, the number of SEOG awards during the period of the audit

E = $\frac{\$50,000}{1460}$ = $34.25 the average acceptable level of error

Z = 1.96 the Z value from Table A.16 associated with a 95 percent confidence level.

Data from the program review was used to estimate $\hat{\sigma}^2$, the population variance. The program review found overawards in the amounts of $1,000, $788, $449, $300, and $1,625, as well as 20 cases with no overawards. From Appendix B, it was found that the sample based estimate of the population variance is:

$$\hat{\sigma}^2 = s^2 = \frac{\Sigma(x_1 - \bar{x})^2}{n - 1}$$

Where:

$$\bar{x} = \frac{\Sigma x_1}{n} = \frac{1000 + 788 + 449 + 300 + 162}{25} = 166.48$$

$$\Sigma(x_1 - \bar{x})^2 = (100 - 166.48)^2 + (788 - 166.48)^2 + (300 - 166.48)^2 + (1625 - 166.48)^2 + 20 \cdot (0 - 166.48)^2 = 3090126.69$$

$$\hat{\sigma}^2 = \frac{\Sigma(x_1 - \bar{x})^2}{n - 1} = 128755$$

Substituting these values into the formula for sample size, we obtain

$$n = \frac{NZ^2 \cdot E^2}{E^2N + Z^2\hat{\sigma}^2}$$

$$= \frac{(1460) \cdot (1.96)^2 \cdot (128755)}{(34.25)^2(1460) + (1.96)^2 (128755)}$$

$$= 327$$

Therefore, the minimum sample size necessary to estimate the total amount of overawards $\pm$ $50,000 is 327. Because the sample size necessary to estimate to number of overawards was larger (386), the University was required to sample a minimum of 386 cases, instead of the proposed sample of 146 cases.

University A conducted the required sample based audit and reported the results in Table 5.1 to the Department.

TABLE 5.1 SEOG AUDIT AT UNIVERSITY A

	Number	Amount
No error found	346	$199,220
Overawards	23	14,623
Missing affidavits	17	10,217
Total	386	224,060

$$\text{Sample Error Rate} = \frac{\text{Number of Overawards (23)} + \text{Number Missing Affadivits (17)}}{\text{Sample Size (386)}}$$

$$= 10.36\%$$

Net SEOG Awards During Period of Audit: $849,481.

Estimated University Liability = $849,481 (Net Awards) X 10.36% (Error Rate) = $88,006

The Department rejected the estimate of total liability because the University employed a faulty computation method. In estimating liability the University had calculated the error rate on the basis of number of errors rather than dollar amount of errors. The correct error rate is:

$$\frac{\text{Amount of Overawards} + \text{Amount of Missing Affidavits}}{\text{Sample Total SEOG}} = 11.09\%$$

Therefore, estimated University liability is:

$849,481 X 11.09% = $94,176.

EXAMPLE 2: BEOG APPLICANT QUALITY CONTROL STUDY

Statistical sampling was employed in a quality control study of BEOG applicants because it offered a wide variety of advantages in the analysis of the universe file of applicants containing over four million records:

- Sampling involved substantial cost savings. In advance, it was estimated that the study would require a minimum of thirty computer reads of the application data. A single computer read of the entire file cost approximately $2,700. Therefore, analysis of the entire universe file would cost a minimum of $81,000. A sample of 20,000 appliations would cost $2,700 to construct. However, after the sample had been drawn, each additional computer read cost only $21 for a total study computer cost of $3,510, a savings of $77,490.
- Sampling introduced only very minor error. For example, in estimating the percent of applicants attending public, 4-year, institutions, the standard error of estimate was less than three-tenths of one percent.
- Sampling speeded the completion of the study. A complete read of the Application File usually requires five hours of computer time and has an average turn-around time of five days using the Department's COMNET facilities. The data file containing a sample of BEOG applications usually only required a few minutes of computer time for each run and had a turn-around time of a few hours.
- Sampling allowed use of a wide range of statistical packages such as SPSS, SAS, OSIRIS and BMDP which are not practical for a file the size of the BEOG Applicant File.
- Sampling made possible a wider range of analyses than would have been possible if the entire file had been utilized. Given the high costs, long time lags, and limited statistical software available, a population-based study could not realistically have explored as many topics as a sample-based study.

EXAMPLE 3: REVIEW OF CWS AUDIT REPORT FROM UNIVERSITY B

The Department received an audit report of CWS awards at University B. Table 5.2 summarizes the audit's findings.

TABLE 5.2 AUDIT REPORT OF CWS AWARDS AT UNIVERSITY B

Total amount of CWS awards during the period of the Audit:		$833,118
Number of CWS Awards:	971	
Simple Random Sample Size:	55	
Number of Overawards Identified:	6	
Error	**Number**	**Amount of Overawards**
Students engaging in profit-making activity for the institution	2	$1,183 742
Students not maintaining satisfactory progress in their course of study	3	300 685 912
Students in default on a NDSL loan	1	1,326
Total	6	$5,148

The Department was concerned with whether the sample provided an adequate basis for projecting total University CWS overawards. To this end, Form B in this manual was used. The completed form with relevant comments is attached.

FORM B

Developing Population Estimates From a Simple Random Sample

STEP		LINE NUMBER
A. Basic Sample Information		
A1 How many cases are in the sample?	$n =$ 55	(1)
A2 How many cases are in the total population from which the sample was drawn?	$N =$ 971	(2)
B. Calculate the sample mean ($\bar{x}$)		
B1 Add the values for all the cases in the sample together	$x_1 + x_2 + x_3 \ldots + x_n = \Sigma x_i =$	(3)
B2 Divide line 3 by line 1	$\Sigma x_i / n = \bar{x} =$ 93.60	(4)
C. Calculate the estimate of the sampling mean standard deviation	$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{s^2}{n} \cdot \frac{N-n}{N}}$	
C1 Using form A, calculate the estimated sample variance	$s^2 =$ 85508.24	(5)
C2 Divide line 5 by line 1	$s^2/n =$ 1554.69	(6)
C3 Subtract line 1 from line 2	$N - n =$ 916	(7)
C4 Divide line 7 by line 2	$\frac{N-n}{N} =$ .9434	(8)
C5 Multiply line 6 by line 8	$\frac{s^2}{n} \cdot \frac{N-n}{N} =$ 1466.63	(9)

In this example we will forego use of Form A and calculate the variance directly:

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

$= ((1,183 - 93.6)^2 +$
$(742 - 93.6)^2 +$
$(300 - 93.6)^2 +$
$(685 - 93.6)^2 +$
$(912 - 93.6)^2 +$
$(1,326 - 93.6)^2 +$
$49 \cdot (0 - 93.6)^2)$
$\div (55 - 1)$

$= (4617445.2) \div (54)$

$= 85508.24$

BEST COPY AVAILABLE

C6 Take the square root of line 9 $\sqrt{\frac{s^2}{n}\cdot\frac{N-n}{N}}$ = 38.30 (10)

The Department used the most commonly accepted confidence level of 95% to estimate total liability.

D. Set the "confidence level" (confidence level is defined on page)

Confidence level = CL = 95 (11)

E. Determine the "Z" or "K" value from the table below

If CL =	If n≥30 then Z =	If n<30 then K =
80%	1.28	2.24
90%	1.64	3.16
95%	1.96	4.44
99%	2.55	10.00

K or Z = 1.96 (12)

F. Calculate the estimate of the population total

$\hat{\tau} = N \cdot \bar{x}$

Multiply line 2 by line 4 $\hat{\tau} = N \cdot \bar{x}$ = $90885 (13)

G. Calculate the confidence interval of the estimated population total

Multiply line 2 by line 12 by line 10

If n≥30 $CI = N \cdot Z \cdot \hat{\sigma}_{\bar{x}}$ = 72886 (14)

OR

If n<30 $CI = N \cdot K \cdot \hat{\sigma}_{\bar{x}}$ = _____ (14)

H. Calculate the upper bound of the confidence interval

$\hat{\tau} + CI$

Add line 14 and 13 $\hat{\tau} + CI$ = 163,771 (15)

I. **Calculate the lower bound of the confidence interval**

$\bar{x}$ – CI

Subtract line 14 from line 13 $\bar{x}$ – CI = $17,999 (16)

J. **Interpreting the results**

Fill in the blanks in the sentence below.

"On the basis of a sample of 55 (line 1) cases, it can be estimated with 95 (line 11) percent confidence, the total value of liability (appropriate name) for the population sampled falls between $17,999 (line 16) and $163,771 (line 15) with the most likely value $91,885 (line 12).

The resulting confidence interval is obviously very wide. The low limit of the interval, $17,999 is less than one-ninth of the high limit of $163,771. Therefore, the Department determined that the audit sample was an insufficient basis for projecting total University liability.

SUMMARY

The three examples contained in this chapter represent only a small fraction of potential statistical sampling applications in OSFA. Nonetheless, taken together, the examples demonstrate that statistical sampling can be very straightforward and need not involve overly complex calculations. The many advantages of statistical sampling can be realized in a great diversity of situations through familiarity with the basic logic of sampling and a few simple formulas.

APPENDIX A

INTRODUCTION TO SAMPLING STATISTICS

The statistics of sampling are presented in several different and largely independent ways. This appendix is a short introduction to, and explanation of, basic sampling statistics. Chapter 4, Computing Sample Statistics, contains a series of forms to aid in calculating a variety of common sample statistics. Appendix B summarizes basic sampling formulas and symbols. Finally, for those who would like a fuller explanation of sampling statistics or more advanced or specialized statistical information, Appendix C presents a short annotated bibliography.

To introduce the statistics of sampling we will consider Artificial University where there were only six student financial aid recipients in 1980. The results of record review that included all six clients are presented in Table A.1.

TABLE A.1: FINANCIAL AID RECIPIENTS AT ARTIFICIAL UNIVERSITY

Student	SEOG Award	SEOG Eligible	NDSL
A	$740	yes	$ 100
B	800	no	1,500
C	800	yes	300
D	672	no	1,500
E	800	no	100
F	700	yes	1,010
Total SEOG Awards		$4,512	
Number of SEOG Overpayments		3	
Total NDSL Awards		$4,510	

The information about Artificial University will provide the basis for our introduction to sampling statistics. First, a set of summary measures for describing population parameters will be presented. Then a series of methods for estimating those population parameters on the basis of data from simple random samples will follow. A word of caution is required here. THE METHODS DESCRIBED BELOW FOR PROJECTING SAMPLE RESULTS TO THE TOTAL POPULATION APPLY ONLY TO SIMPLE RANDOM SAMPLES AND SYSTEMATIC SAMPLES. Other sample designs, such as stratified cluster or discovery sampling employ different formulas.

TOTAL, POPULATION SIZE, MEAN, VARIANCE, STANDARD DEVIATION, AND DISTRIBUTION

A primary use of statistics is to summarize complex data into a few simple measures. The first step in summarizing data is to note how many cases are in the population under review. The number of cases in the population is usually symbolized by a capital 'N'. For Artificial University (AU), N=6, since there are 6 financial aid recipients. A

second common summary measure is the population total. The population total is usually symbolized by the Greek letter τ (pronounced tou). The population total, τ, is calculated by summing all the values of all the individual cases. Individual case values are symbolized by 'x_i'. The operation of summing all the cases in the population can be symbolized by

$$\Sigma x_i$$

Σ, the large Greek letter sigma means "the sum of". Therefore, Σx_i means "the sum of all individual cases,"

$$\Sigma x_i = x_1 + x_2 + x_3 \ldots\ldots x_N \text{ (the Nth, or last case in the population)}$$

In AU the total value of SEOG awards is:

$$\tau = \Sigma x_i = 740 + 800 + 800 + 672 + 800 + 700 = 4512$$

The number of cases in the population and the population total can be combined to produce a third common summary measure; the mean or average. The population mean is symbolized by the small Greek letter 'μ' (pronounced 'm/y/oo') and is calculated by dividing the population total (τ) by the number of cases in the population (N), thus:

$$\mu = \frac{\tau}{N} = \frac{\Sigma x_i}{N} \qquad \text{(A.1)}$$

For AU, the mean value of SEOG awards is:

$$\mu = \frac{\tau}{N} = \frac{4512}{6} = \$752$$

Table A.2 summarizes SEOG awards and NDLS loans at Artificial University in terms of number of cases, total and mean.

TABLE A.2: SEOG AWARD AND NDSL AT ARTIFICIAL UNIVERSITY

	SEOG	NDSL
Number of Cases (N)	6	6
Total Awards (τ)	4512	4510
Mean Award (μ)	752	751.67

Number of cases in a population, population total and population mean are generally not, by themselves, sufficient to describe and adequately summarize the data under study. Table A.2 reveals practically no difference between these summary measures describing SEOG and NDSL. However, returning to Table A.1, we can see that all SEOG award amounts are clustered between $672 and $800 whereas NDSL award amounts are much more variable, ranging from $100 to $1,500. To represent this important difference, a measure of dispersion (or variability or spread) is also needed. As the words "dispersion", "variability", and "spread" suggest, the measures that summarize this characteristic indicate the extent to which individual cases are scattered about the mean.

The two most common measures of dispersion employed in statistics are 'variance' and 'standard deviation'. The variance of a population is represented by the symbol 'σ^2' (small sigma squared). The variance of a population is calculated by the formula:

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} \qquad (A.2)$$

Where:

σ^2 = the variance

x_i - values of the individual cases

μ = the mean value of the cases

N = the number of cases in the population

The steps involved in the calculation of the variance are:

1. The total value is computed (τ).
2. The mean value of the cases is computed (μ).
3. The deviations of the individual award amounts from the mean are computed ($x_i - \mu$).
4. The deviations are squared then totaled ($\Sigma(x_i - \mu)^2$).
5. The sum of the squared deviations is divided by the number of cases, 'N'.

Table A.3 illustrates calculation of the variance of SEOG awards in AU.

TABLE A.3: COMPUTATION OF THE VARIANCE

1 Student	2 SEOG Award(x_i)	3 Mean Award (μ)	4 ($x_i - \mu$)	5 $(x_i - \mu)^2$
A	740	752	-12	144
B	800	752	48	2304
C	800	752	48	2304
D	672	752	-80	6400
E	800	752	48	2304
F	700	752	-52	2704
Total =	= 4512			$\Sigma (X_i - \mu)^2 = 10400$

$N = 6$

$\tau = \Sigma x_i = 4512$

$\mu = \frac{\tau}{N} = 752$

$\sigma^2 = \frac{\Sigma (x_i - \mu)^2}{N}$

$= \frac{10400}{6} = 1733$

The variance of SEOG awards at AU is 1733. This value represents the average variability of squared dollars. To obtain a measure of dispersion expressed in terms of the original values, we calculate the standard deviation. The standard deviation, which is symbolized by the small Greek letter 'σ', is the square root of the variance. The formula for the standard deviation is:

$$\sigma = \sqrt{\frac{\Sigma (x_i - \mu)^2}{N}} \qquad (A.3)$$

The standard deviation of SEOG awards at AU is $41.60. The much greater dispersion of NDSL at AU is represented by a standard deviation of $611.

ATTRIBUTES

To this point the discussion has focused exclusively on continuous variables such as dollar amount of SEOG awards. However, statistics can also be applied to categorical attributes such as program eligibility, which have no natural numeric values attached to them. The question therefore arises as to how to calculate totals, means, standard deviations, etc. for case attributes. To give categories a mathematical representation, cases in the category of interest are commonly assigned a value of 1 and all other cases are assigned a value of 0.

Categorical case attributes can be summarized in terms of frequency and proportion. At Artificial University, 3 students, or .5 of all financial aid recipients, are SEOG eligible. Population frequency will be symbolized as a large 'F' and proportion of the population having a certain attribute by a large 'P'. Therefore, for SEOG eligibility at AU, F=3 and P=.5. Either frequency or proportion can be used to calculate population mean, variance and standard deviation:

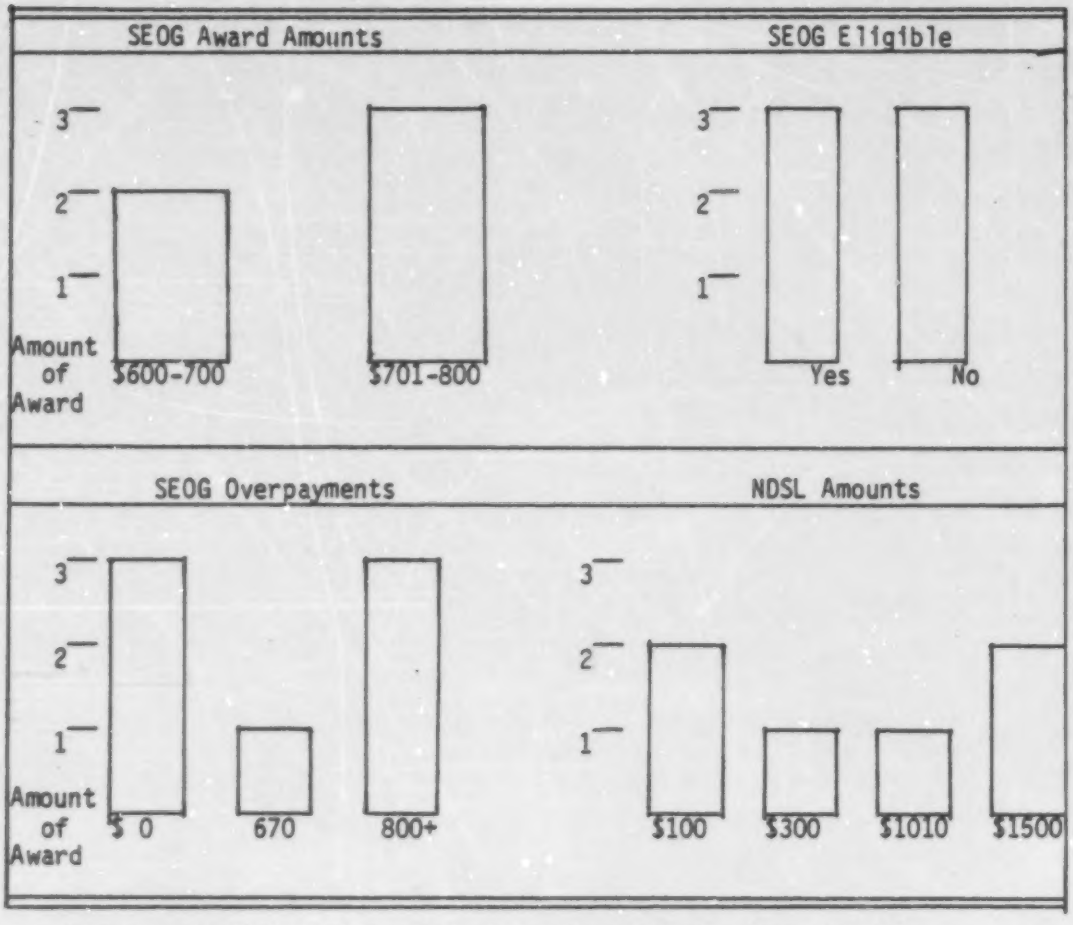
Frequency	Proportion	
$\tau = F$	$\tau = P \cdot N$	(A.4)
$\mu = F/N$	$\mu = P$	(A.5)
$\sigma^2 = \frac{F - F^2/N}{N}$	$\sigma^2 = P\,(1 - P)$	(A.6)
$\sigma = \sqrt{\frac{F - F^2/N}{N}}$	$\sigma = \sqrt{P\,(1 - P)}$	(A.7)

For SEOG eligibility at AU, F = 3, P = .5, ^{2} = .25 and = .5.

One additional way data can be summarized is to graph its frequency distribution. Table A.4 presents graphs of frequency distributions of the data contained in Table A.1. As Table A.4 shows, frequency distributions can be shaped in many different ways.

For reasons that will become clear below, sampling statistics make frequent use of one particular type or shape of frequency distribution; the normal distribution.

TABLE A.4: FREQUENCY DISTRIBUTION OF STUDENT FINANCIAL AID RECIPIENTS AT ARTIFICIAL UNIVERSITY



The normal distribution is a frequency distribution which is bell-shaped. Table A.5 gives an example of an approximately normal distribution. The results of a sample of the SAT math scores for 10,000 high school seniors are graphed in terms of frequency of test score.

Relative frequency of occurrence in a normal distribution is governed by distance from the mean measured in standard deviations. In Table A.5 68 percent of the cases fall within one standard deviation of the average score of 500. Because, for the SAT math scores, the standard deviation is 100 points, approximately 68 percent of the scores fall between 400 and 600. Similarly, approximately 95.4 percent of the cases fall within two standard deviations of the mean and 99.7 percent of the cases fall

within three standard deviations. What is true of SAT math scores is true of any normally distributed variable. In any real situation a distribution, at best, will be only approximately normally distributed. However, in many situations, the approximation is very close. Table A.6 gives the percent of cases in terms of distance from the mean for normal distribution.

TABLE A.6: NORMAL DISTRIBUTION

Percent of Cases	Distance from the Mean Measured in Standard Deviations (Z values)
50.00	.67
60.00	.84
70.00	1.04
80.00	1.28
90.00	1.65
95.00	1.96
98.00	2.33
99.00	2.57
99.90	3.30
99.99	3.90

To determine the range around the mean that contains a certain pre-specified percent of cases, we use the following formula:

$$\mu \pm Z\sigma \qquad \text{(A.8)}$$

For example, if we wanted to know the range around the mean that contained 95 percent of the SAT scores we would look up the Z value corresponding to 95 percent, which is 1.96. We know that standard deviation of SAT scores is 100 and the mean is 500. Substituting these values into equation 4.8 we obtain:

$$\mu \pm Z\sigma$$
$$= 500 \pm 1.96 \times 100$$
$$= 500 \pm 196$$
$$= 304 \text{ to } 696$$

Therefore, 95 percent of the SAT scores fall between 304 and 696.

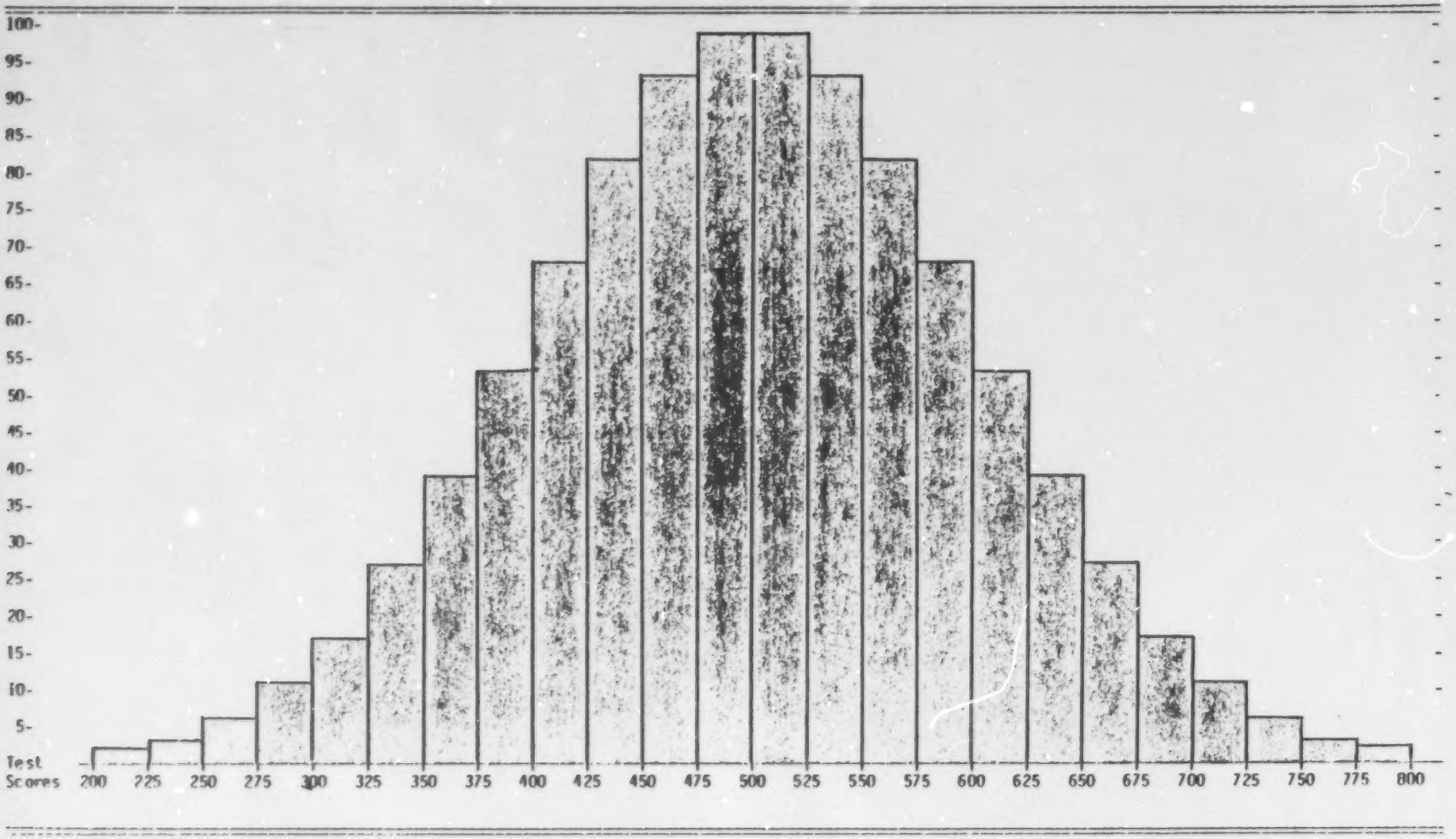
POPULATION AND SAMPLE SYMBOLS

To clearly distinguish between summary measures which describe a population and those that describe a sample, different sets of symbols are used to represent the two sets of measures. As indicated earlier, a large N is used to symbolize the size of the population. A small n is used to symbolize the size of a sample. When a sample statistic is used to estimate a population parameter a '^' is placed over the symbol to indicate that it is an estimate. For example, $\hat{\tau}$, is the symbol for an estimate of the population total. Table A.7 summarizes the symbolism used in sampling statistics.

TABLE A.7: SAMPLING SYMBOLS

Summary Measure	Population Symbol	Sample Symbol
Number of Cases	N	n
Total	τ	$\hat{\tau}$
Mean (average)	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s
Frequency	F	f
Proportion	P	p

TABLE A.5: EXAMPLE OF NORMAL DISTRIBUTION (SAT Math Scores for a Sample of 10,000 High School Seniors)



A10

BEST COPY AVAILABLE

ESTIMATING THE POPULATION MEAN

Although we know the mean value of SEOG grants at AU we will act as though this value is unknown to us and will estimate it through simple random sampling. We begin with a sample size of 2. The number of cases in a sample is represented by a small 'n'. In the population of 6 SEOG recipients at AU, there are 15 possible simple random samples, without replacement, of 2 cases each. Table A.8 lists all possible samples of two students in Column 2. The mean value of SEOG for each sample is listed in column 3.

TABLE A.8: SAMPLES OF SEOG RECIPIENTS AT AU
(n=2)

(1) Sample Number	(2) Students	(3) Sample Mean ($\bar{x}$)	(4) Population Total Estimate ($\bar{x} \cdot N$)	(5) Error of Estimate	(6) Squared Error of Estimate
1	AB	770	4620	18	324
2	AC	770	4620	18	324
3	AD	706	4236	-46	2116
4	AE	770	4620	18	324
5	AF	720	4320	-32	1024
6	BC	800	4800	48	2304
7	BD	736	4416	-16	256
8	BE	800	4800	48	2304
9	BF	750	4500	- 2	21
10	CD	736	4416	-16	256
11	CE	800	4800	48	2304
12	CF	750	4500	- 2	4
13	DE	736	4416	-16	256
14	DF	686	4116	-66	4356
15	EF	750	4500	- 2	4
AVERAGE		752	4512	0	1077

If we examine the 15 possible samples listed in Table A.8, we see variation in the results. Sample 6, for instance, has an average of 800 whereas sample 14 has an average of 686. However, computing the average of all the sample means we get a value of $752; the exact value of the mean of population. This fact is of great importance because it

demonstrates that simple random sampling will on the average produce a sample mean ($\bar{x}$) which is equal to the population mean (μ). Therefore we can conclude that the sample mean is an unbiased estimate of the population mean. ('Bias' is defined on page 11).

ESTIMATING THE POPULATION TOTAL AND STANDARD DEVIATION

The sample mean can be used to calculate an estimate of the population total (column 4, Table A.8). The appropriate formula is:

$$\hat{\tau} = N \cdot \bar{x} \qquad (A.9)$$

The '^' over the 'τ' indicates that it is an estimate of the population total. Because $\bar{x}$ is an unbiased estimation of the population mean, $N \cdot \bar{x}$ is an unbiased estimator of the population total. This fact is illustrated in column 4 of Table A.8. An estimate of the population total is calculated on the basis of each of the 15 samples. The mean value of the population total estimates is \$4512, the exact value of the true population total.

The sample-based estimate of the standard deviation of the population is symbolized by a small 's'. The formula for a continuous variable is,

$$s = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n - 1}} \qquad (A.10)$$

Where:

x_i is an individual case value in the sample

$\bar{x}$ is the mean of the sample

n is the number of cases in the sample

The formula for s for a categorical variable is,

$$s = \sqrt{\frac{f - f^{2}/n}{n - 1}} \qquad (A.11)$$

Where:

f is the frequency of the category of interest

The equations for s are identical to the equations for σ with the exception of the -1 in the denominator. The standard deviation of a sample is, on the average, less than the standard deviation of the population and is therefore a biased estimator of the population standard deviation without the corrective factor of reducing 'n' by one.

STANDARD ERROR OF THE MEAN

Although the sample mean is an unbiased estimator of the population mean, as Table A.8 illustrates, there can be great dispersion of sample means. One measure of the accuracy of the sampling plan is the mean square error (MSE) of the estimates of the mean.

$$MSE = (\text{error of estimate})^2/(\text{number of samples}) \qquad (A.12)$$

Returning to the data in Table A.8, the error of estimate for each sample can be found in column 5 and the squared error of estimate in column 6. Substituting these values into equation A.12 we obtain:

$$MSE = 16160/15 = 1077$$

To state the error of the estimate in terms of dollars rather than squared dollars we take the square root of the MSE to produce the standard error of the mean; $1077 = \$32.82$. The standard error of the mean is a measure of the reliability or precision of a sampling plan. The standard error of the mean is the standard deviation of sample means and symbolized by '$\sigma_{\bar{x}}$'.

In actual practice we almost never have the data necessary to directly calculate the standard error of the mean for a sampling procedure. We usually do not know the true population mean and draw only one, not fifteen, samples. Therefore, an alternative method of determining the reliability of a sampling procedure is needed.

Fortunately, the standard error of the mean can be estimated on the basis of data from a single sample. For simple random sampling from a finite population without replacement the formula for the estimate of $\hat{\sigma}_{\bar{x}}$ is:

$$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N}} \qquad (A.13)$$

Substituting the equation A.10 for 's' into the equation we obtain:

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} \cdot \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N}} \qquad (A.14)$$

$$= \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n(n-1)} \cdot \frac{N-n}{N}} \qquad (A.15)$$

When sampling from an infinite population or sampling with replacement the formula for $\hat{\sigma}_{\bar{x}}$ can be simplified to:

$$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n(n-1)}} \qquad (A.16)$$

For a sufficiently large sample ($n \geq 30$), $\hat{\sigma}_{\bar{x}}$ will be approximately normally distributed with mean of μ. This mathematical fact, known as the Central Limit Theorem, is significant because it allows calculation of confidence intervals on the basis of known characteristics of the normal distribution.

We know, from Table A.6, that 95 percent of the cases fall within 1.96 standard deviations of the distribution mean. Because the central limit theorem states that the mean of the $\hat{\sigma}_{\bar{x}}$ distribution is μ, we can conclude that 95 percent of all sample means, x_i will fall within 1.96 $\hat{\sigma}_{\bar{x}}$ of the true population mean, μ. In other words, if 1000 samples of the same size were drawn from a single population, approximately 950 of the sample means would fall within 1.96 $\hat{\sigma}_{\bar{x}}$ of the population mean.

If 95 percent of the possible values of $\bar{x}$ fall within 1.96 $\hat{\sigma}_{\bar{x}}$ of μ, then μ will not be further than 1.96 $\hat{\sigma}_{\bar{x}}$ from 95 percent of the possible values of $\bar{x}$. This leads us to the final step in our reasoning, the pay-off: If we estimate a confidence interval of $x \pm 1.96\ \hat{\sigma}_{\bar{x}}$ and if we construct a large number of such intervals, 95 percent of the interval estimates will include . Therefore the commonly used phrase: "at 95% confidence."

CONFIDENCE INTERVAL FOR THE MEAN

The formula for calculating the confidence interval for the mean is:

$$CI_{\bar{x}} = \bar{x} \pm Z \cdot \hat{\sigma}_{\bar{x}} \qquad (A.17)$$

If we substitute equation A.15 for $\hat{\sigma}_{\bar{x}}$ into equation 4.17 we obtain:

$$CI_{\bar{x}} = \bar{x} \pm Z \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n\,(n-1)} \cdot \frac{N-n}{N}} \qquad (A.18)$$

The steps involved in the calculation of the confidence interval are:

1. The sample mean ($\bar{x}$) is computed
2. The standard error of the mean is computed by:
   - 2A The sum squared deviations around the mean is computed $\left(\Sigma(x_i - \bar{x})^2 \right)$
   - 2B The demonimator, n(n-1) is computed
   - 2C The "finite population correction" $\frac{N - n}{N}$, is computed
   - 2D The results of steps 2A, 2B, and 2C are substituted into formula for A.15 and the square root taken.
3. The Z value is obtained from Table A.6.
4. The sample mean ($\bar{x}$), the Z value, and the standard error of mean are substituted into the equation for the confidence interval.

Returning to Artificial University records, we draw a three-student simple random sample of SEOG recipients to estimate the average grant amount of the population with a 95 percent confidence level. From Table A.1 the students selected are A, C, and D. Because we have selected a 95 percent confidence level, from Table A.6, Z=1.96, Table A.9 illustrates the calculation of the confidence interval.

TABLE A.9: EXAMPLE OF CALCULATING A CONFIDENCE INTERVAL

Sample Students	SEOG Awards
A	740
C	800
D	672

Step

1. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{740 + 800 + 700}{3} = \frac{2240}{3} = 746.67$

2. $\hat{\sigma}_{\bar{x}} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n(n-1)} \cdot \frac{N-n}{N}}$

2.A $\Sigma(x_i - \bar{x})^2 = (740 - 746.67)^2 + (800 - 746.67)^2 + (672 - 746.67)^2$

$= 44.44 + 284.44 + 5575.11 = 8464$

2.B $n(n-1) = 3(3-1) = 3(2) = 6$

2.C $\frac{N-n}{N} = \frac{6-3}{6} = \frac{3}{6} = .5$

2.D $\hat{\sigma}_{\bar{x}} = \sqrt{\frac{8464}{6} \cdot .5} = \sqrt{1410.67 \cdot .5} = \sqrt{705.33} = 26.56$

3. $Z = 1.96$ (95% confidence level)

4. $CI_{\bar{x}} = \bar{x} \pm Z \cdot \hat{\sigma}_{\bar{x}}$

$= 746.67 \pm 1.96 \cdot 26.56 = 746.67 \pm 52.05$

$= 693.95$ to 798.72

SMALL SAMPLES

The confidence interval obtained does, indeed, include the true population mean value of $752. This result, however, must be attributed to good luck rather than good statistics. As already indicated, the estimation equations that were employed assume a sample size of at least thirty cases. For sample sizes under thirty, the Central Limit Theorem is not generally applicable. Tchebysheff's Theorem can be used as an alternative to the Central Limit Theorem for making population estimates on the basis of a small sample. Tchebysheff's Theorem states that at least $(1-1/K^2)$ of a set of measurements will lie within K standard deviations of their mean. To employ Tchebysheff's Theorem, simply replace the 'Z' value in equation A.17 with a 'K' value from Table A.10. Thus, for a small sample, the equation for estimating the confidence interval becomes:

$$CI_{\bar{x}} = \bar{x} \pm K \cdot \hat{\sigma}_{\bar{x}} \qquad (A.19)$$

To calculate the correct confidence interval for the sample of three AU SEOG recipients, the same steps as before are performed, except in this case Table A.10 is used to obtain a 'K' value rather than Table A.6 to obtain a 'Z' value. The results of the computations are:

$$\begin{aligned} CI_{\bar{x}} &= \bar{x} \pm K \cdot \hat{\sigma}_{\bar{x}} \\ &= \$746.67 \pm \$118.72 \text{ or} \\ &= \$865.39 \text{ to } \$627.95 \end{aligned}$$

The results can be described as follows:

"On the basis of a three-student simple random sample of Artificial University SEOG recipients, it can be estimated, with 95 percent confidence, the mean value of SEOG awards at AU falls between $627.95 and $865.39 with the most likely value $746.67." "95 percent confidence" means that if we were to follow the same procedure for drawing multiple three-student samples of AU SEOG recipients, 95 percent of the resulting confidence intervals would contain the true population mean. Table A.11 verifies this result. Confidence interval estimates for average SEOG

awards at AU based on all possible three-student samples at 95 percent confidence are displayed. Of the 20 confidence intervals 19, or 95 percent contain the true mean of $752.

TABLE A.10: TCHEBYSHEFF'S THEOREM

Percent of Cases	Distance From the Mean (Measured in Standard Deviations) (K Values)
50.00	1.41
60.00	1.58
70.00	1.83
80.00	2.24
90.00	3.16
95.00	4.47
98.00	7.07
99.00	10.00
99.90	31.62
99.99	100.00

CONFIDENCE INTERVAL FOR THE TOTAL

The confidence interval for a sample-based estimate of the population total is obtained by is simply multiplying the confidence interval for the mean $CI_{\bar{x}}$ by N, the number of cases in the population:

$$CI = N \cdot CI_{\bar{x}} \tag{A.20}$$

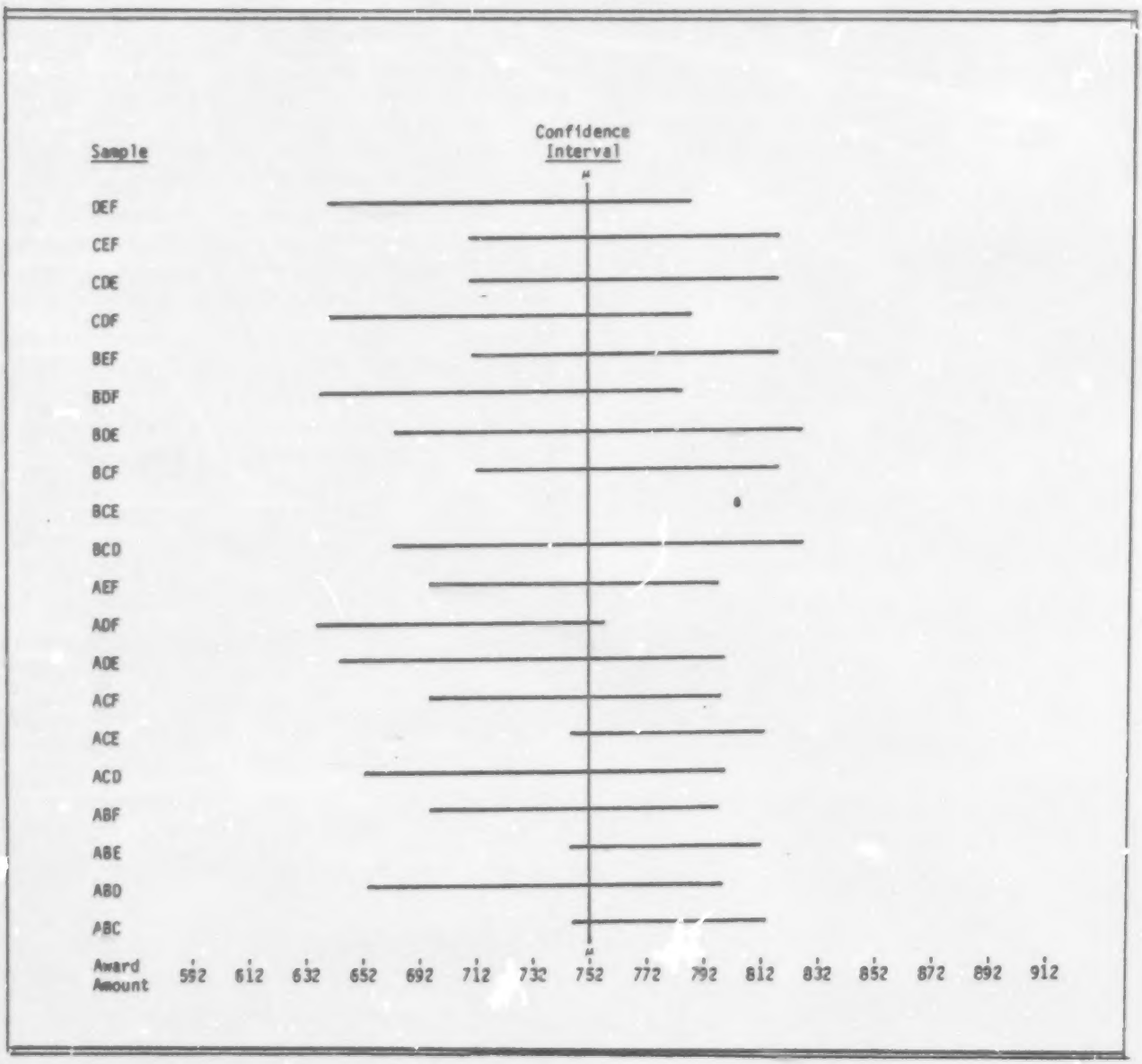
For $n \geq 30$:

$$CI = N \cdot (\bar{x} \pm Z\,\hat{\sigma}_{\bar{x}}) = N\bar{x} \pm N \cdot Z\hat{\sigma}_{\bar{x}} = N\bar{x} \pm N \cdot Z \cdot \hat{\sigma}_{\bar{x}} \tag{A.21}$$

For $n < 30$:

$$CI = N \cdot (\bar{x} \pm K\hat{\sigma}_{\bar{x}}) \qquad CI = N\bar{x} \pm N \cdot K \cdot \hat{\sigma}_{\bar{x}} \tag{A.22}$$

TABLE A.11: EXAMPLE OF RELATION BETWEEN SAMPLE CONFIDENCE INTERVAL AND POPULATION MEAN (Confidence interval estimates of average SEOG awards at AU based on three-case samples at 95% confidence)



THE RELATION BETWEEN CONFIDENCE INTERVAL AND CONFIDENCE LEVEL

In the equation for calculating confidence intervals for the mean, $CI=\bar{x} \pm Z\sigma_{\bar{x}}$ (Formula A.17), a direct relationship can be seen between the confidence level (as represented by 'Z') and the confidence interval (CI). The higher the 'Z' value the wider the confidence interval. This is the result of the commonsensical fact that the more certain we wish to be that the true population mean falls somewhere in the confidence interval, the wider the interval must be. Table A.12 illustrates the relationship between confidence intervals and confidence levels. For the example given in the table, with a sample size of 250 at a 95 percent confidence level, the confidence interval is $\pm$ 6.2 percent. With a higher confidence level of 99 percent, the confidence interval grows to $\pm$ 8.1 percent.

MINIMUM NECESSARY SAMPLE SIZE

For a simple random sample drawn without replacement the formula for determining minimum necessary sample size is

$$n = \frac{N \cdot Z^2 \cdot \hat{\sigma}^2}{E^2 \cdot N + Z \cdot \hat{\sigma}^2} \qquad \text{(A.23)}$$

Where:

n = minimum necessary sample size

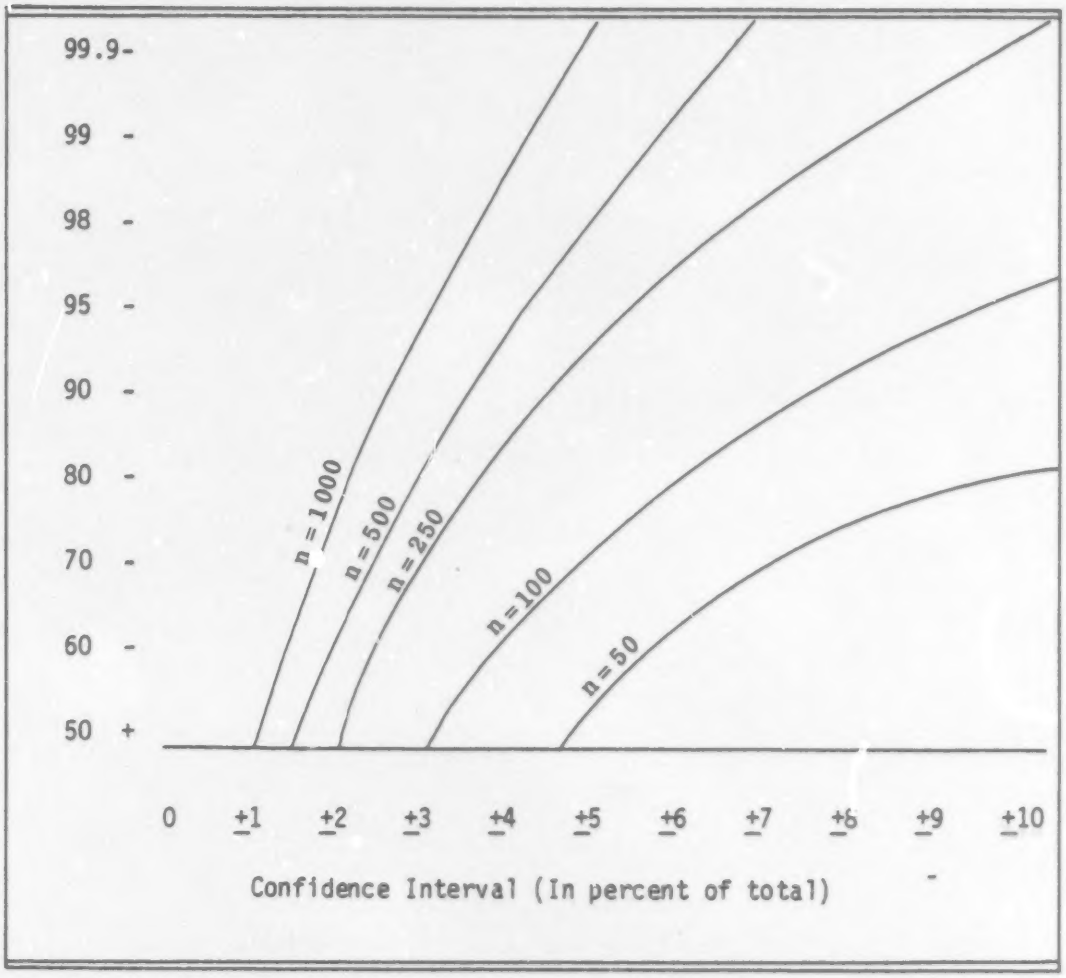
Z = Z value based on desired confidence level. (See Table A.7 to obtain appropriate value.)

E = Acceptable average level of error of estimate (Confidence Interval)

$\hat{\sigma}^2$ = Estimate of population variance.

As an illustration, consider a quality control review of BEOG applications. The reviewers wished to establish, at the 90 percent confidence level, the average family income of applicants within a $300 confidence interval. The sample was to be drawn from a data file which contained approximately 6,000,000 applications. From previous studies, the reviewers estimated the family income standard deviation at $9,000. Translating these facts into the proper statistical notation:

TABLE A.12: EXAMPLES OF THE RELATION BETWEEN CONFIDENCE INTERVALS AND CONFIDENCE LEVELS



(Confidence level by confidence interval for various sample sizes for a two-category variable with a 50/50 population distribution based on simple random sampling with replacement from a large population.)

$N = 6{,}000{,}000$ (The number of cases in the population sampled)

$E = 300$ (The acceptable average level of error of estimate)

$Z = 1.65$ (The Z value associated with a 90% confidence level from Table A.7)

$\hat{\sigma}^2 = (9{,}000)^2 = 81{,}000{,}000$ (Estimated variance of family income)

Substituting these values into equation A.19, we obtain:

$$n = \frac{(6{,}000{,}000) \cdot (1.65)^2 \cdot (9{,}000)^2}{(300)^2 \cdot (6{,}000{,}000) + (1.65) \cdot (9{,}000)^2}$$

$$= 2450$$

THE RELATION BETWEEN SAMPLE SIZE AND POPULATION SIZE

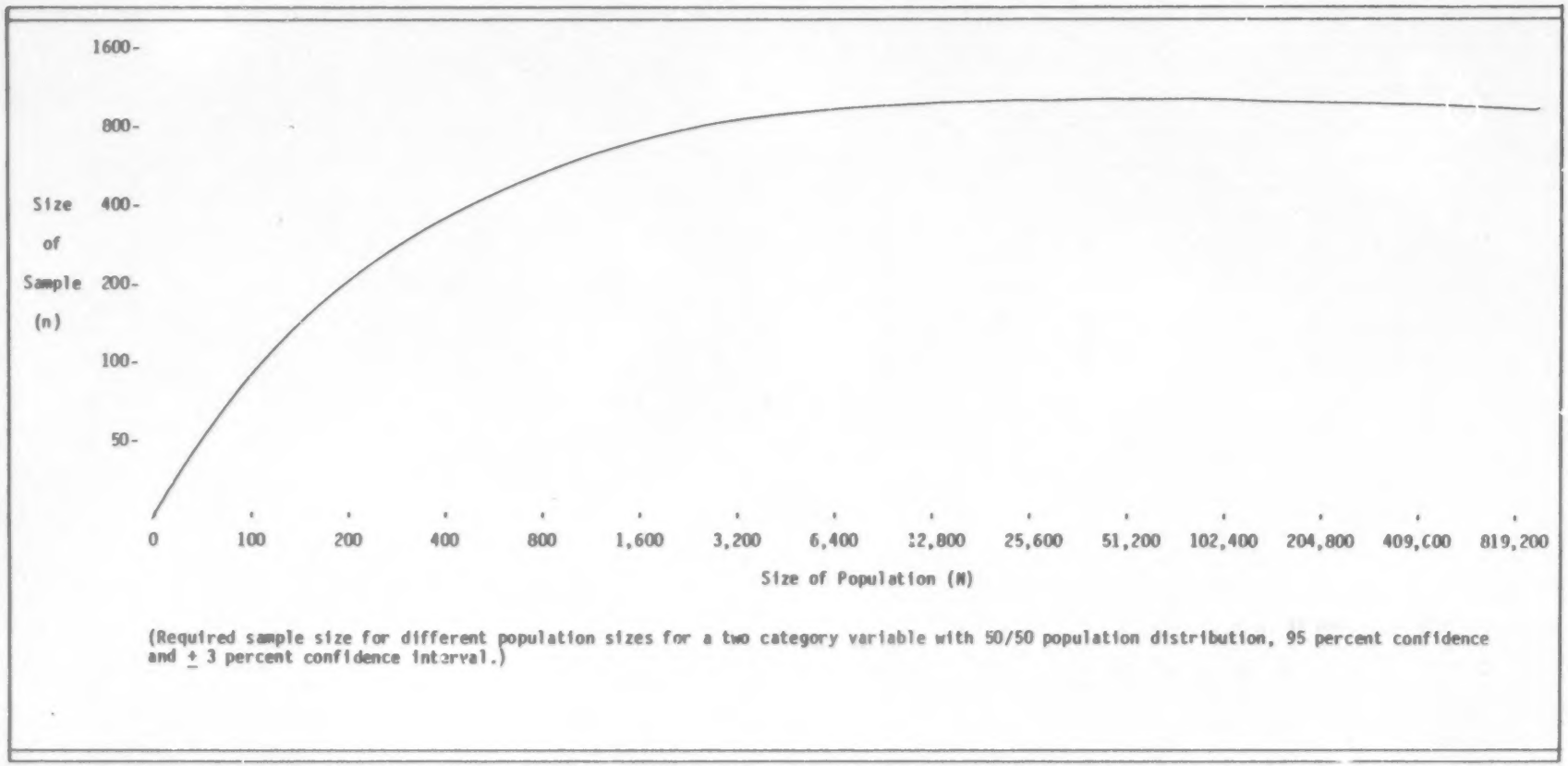
Examination of equation A.23 for minimum necessary sample size reveals that there is no simple, direct relation between population size and necessary sample size. Therefore, it is not possible to determine necessary sample size as a simple percent of the population. An illustration of the complex relation betwen sample and population size is contained in Table A.13. In the case illustrated, for populations under 3,200 cases, the necessary sample size is directly related to population size. However, for large populations the required sample size is almost completely independent of the size of the population. Thus, in the example, the sample size required for a population of 25,000 is almost equal to the sample size for a population of a 1,000,000.

This result is of great importance. In a large population, the necessary sample size depends primarily on the variability of the population and only a little on the fraction of the population sample. Many people intuitively feel, as an example, that a 30% sample of a population of 200 would yield much more precise results than a .1% sample of a population of a million. In fact, as Table A.13 demonstrates, the exact opposite is true. This helps explain the great costs savings that are possible using sampling with a large population.

PRACTICAL PROBLEMS IN DETERMINING SAMPLE SIZE

In many practical circumstances, all the information needed to calculate minimum necessary sample size is not readily available. When,

TABLE 4.13: EXAMPLE OF RELATION BETWEEN MINIMUM NECESSARY SAMPLE SIZE AND POPULATION SIZE



(Required sample size for different population sizes for a two category variable with 50/50 population distribution, 95 percent confidence and ± 3 percent confidence interval.)

A23

BEST COPY AVAILABLE

N, the number of cases in the population, is unknown, or when N is very large as in the example above, the formula for minimum necessary sample size can be simplified to:

$$n = \frac{Z^2 \cdot \hat{\sigma}^2}{E^2} \qquad (A.24)$$

Returning to the previous example, and substituting in the appropriate values, we get:

$$n = \frac{(1.65)^2 \cdot (9{,}000)^2}{(300)^2} = 2450$$

The solutions to necessary sample size were identical for the two formulas. The results given by equation A.23 and equation A.24 will diverge significantly only when the sample size is 5% or greater of the total population.

A common situation is that the variance of the population to be sampled is unknown. When this is the case, there are a variety of ways of estimating $\hat{\sigma}^2$, the population variance:

- $\hat{\sigma}^2$ can be estimated on the basis of previous studies or past experience
- A small pilot sample can be drawn to estimate $\hat{\sigma}^2$.
- If sampling from an approximately normally distributed population, $\hat{\sigma}^2$ can be roughly estimated as

$$\hat{\sigma}^2 = \frac{R^2}{25}$$

  Where R is the range. The range is the highest known value minus the lowest known value.

- Where the variable being studied is categorical, $\hat{\sigma}^2$ can be estimated by assuming the maximum variance and setting $\hat{\sigma}^2 = .25$.

A third common problem in determining necessary sample size arises in situations where the acceptable error is defined in terms of estimating the population total rather than estimating the population mean. An example would be to estimate total aid overpayments (made by an institution) rather than average overpayment. Suppose an auditor wanted to draw a sample which would allow estimation of total SEOG overpayments (within a margin of error of $\pm$ \$100,000) at a University having 7,600 SEOG recipients. To determine required sample size it would first be necessary to convert the total acceptable error into average acceptable error by dividing the total acceptable error (\$100,000) by the number of cases in the population. (N = 7,600 SEOG recipients).

Thus:

$$E = (\text{Total acceptable error})/N = (\text{Average acceptable error})$$

$$E = (\$100,000)/(7,600) = \$13.16 \qquad \text{(A.25)}$$

The resulting value can then be employed in equation 4.23 to establish minimum necessary sample size.

SELF-TESTING REVIEW

Based on the data in Table A.12, complete the following exercises. Answers can be found on the following page.

TABLE A.14: CWS EARNINGS AT ARTIFICIAL UNIVERSITY

Student	CWS Earnings
A	\$ 470
B	750
C	1,100
D	120
E	590
F	600

1. Compute the population size (N), mean (u), total (τ), variance (σ^2), and standard deviation (σ).

2. On the basis of a sample of students C, B, and E, calculate the sample mean ($\bar{x}$), total ($\hat{\tau}$), variance (s^2) and standard deviation (s).

3. On the basis of a sample of students D, E, A, C, estimate the population total ($\hat{\tau}$) and a confidence interval with an 80 percent confidence level.
4. For a university with 375 CWS recipients, what is the necessary sample size to estimate the population mean $\pm$ \$60 at a 90 percent confidence level assuming a population standard deviation of N ~ \$294.32?

ANSWERS TO SELF-TESTING REVIEW

1. $N = 6$

$$\mu = \frac{\Sigma x_i}{N} = \frac{3630}{6} = \$605$$

$$\tau = \Sigma x_i = \$3630$$

$$\sigma^2 = \frac{\Sigma (x_i - \bar{x})^2}{N} = \frac{519750}{6} = 86625$$

$$\sigma = 294.32$$

2. $$\bar{x} = \frac{\Sigma x_i}{n} = 813.33$$

$$\hat{\tau} = \Sigma x_i = 2440$$

$$s^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1} = \frac{136066.67}{2} = 68033$$

$$s = 260.83$$

3. $CI = N\bar{x} \pm N \cdot K \cdot \sigma_{\bar{x}}$

At a confidence level of 80%, K=2.24

$N = 6$

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{2280}{4} = 570$$

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n(n-1)} \cdot \frac{N-n}{N}}$$

$$= \sqrt{\frac{493800}{12} \cdot \frac{2}{6}}$$

$$= 117.12$$

$CI = 6 \times 570 \pm 6 \times 2.24 \times 117.12$

$= 3420 \pm 1574.07$

= \$1845.93 to \$4994.07

4. $$n = \frac{NZ^2 \cdot \hat{\sigma}^2}{E^2N + Z^2\hat{\sigma}^2}$$

where:

$N = 375$

$Z = 1.65$ at 90% confidence

$E = 60$

$\hat{\sigma} = 294.32$

$$n = \frac{(375) \cdot (1.65)^2 \cdot (294.32)^2}{(60)^2 \cdot (375) + (1.65) \cdot (294.32)^2}$$

$= 56$

BEST COPY AVAILABLE

APPENDIX H: SAMPLING SYMBOLS AND FORMULAS
(For Simple Random Sampling Without Replacement)

	Population		Sample	
	Continuous Variable	Categorical Variable	Continuous Variable	Categorical Variable
Number of Cases	N	N	n	n
Population Frequency of Attribute of Interest		F		$\hat{F} = f \cdot N/n$
Population Proportion of Attribute of Interest		$P = F/N$		$\hat{P} = p = f/n$
Population Total	$\tau = \Sigma x_i$	$\tau = F$	$\hat{\tau} = \bar{x} \cdot N$	$\hat{\tau} = p \cdot N = \bar{x} \cdot N$
Population Mean	$\mu = \Sigma x_i/N$	$\mu = F/N = P$	$\hat{\mu} = \bar{x}$ $\bar{x} = \Sigma x_i/n$	$\hat{\mu} = \bar{x}$ $\bar{x} = f/n = p$
Population Standard Deviation	$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}} = \sqrt{\frac{\Sigma x_i^2 - (\Sigma x_i)^2/N}{N}}$	$\sigma = \sqrt{\frac{F - F^2/N}{N}} = \sqrt{P(1-P)}$	$\hat{\sigma} = s$ $s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\Sigma x_i^2 - (\Sigma x_i)^2/n}{n-1}}$	$\hat{\sigma} = s$ $s = \sqrt{\frac{f - f^2/n}{n-1}}$

BEST COPY AVAILABLE

BEST COPY AVAILABLE

APPENDIX B: SAMPLING SYMBOLS AND FORMULAS (Continued)
(For Simple Random Sampling Without Replacement)

	Population		Sample	
	Continuous Variable	Categorical Variable	Continuous Variable	Categorical Variable
Standard Error of the Mean			$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N}}$ $= \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n(n-1)} \cdot \frac{N-n}{N}}$	$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N}}$ $= \sqrt{\frac{f - f^2/n}{n(n-1)} \cdot \frac{N-n}{N}}$
Confidence Interval for Estimate and Population Mean ($n \geq 30$)			$CI_{\bar{x}} = \bar{x} \pm Z\sigma_{\bar{x}}$ (See Table 4.6 for Z values)	$CI_{\bar{x}} = \bar{x} \pm Z\sigma_{\bar{x}}$ (See Table 4.6 for Z values)
Confidence Interval for Estimate of Population Mean ($n < 30$)			$CI_{\bar{x}} = \bar{x} \pm K\sigma_{\bar{x}}$ (See Table 4.10 for K values)	$CI_{\bar{x}} = \bar{x} \pm K\sigma_{\bar{x}}$ (See Table 4.10 for K values)
Confidence Interval for Estimate of Population Total ($n \geq 30$)			$CI_t = t \pm N \cdot Z \cdot \sigma_{\bar{x}}$	$CI_t = t \pm N \cdot Z \cdot \sigma_{\bar{x}}$
Confidence Interval for Estimate of Population Total ($n < 30$)			$CI_t = t \pm N \cdot K \cdot \sigma_{\bar{x}}$	$CI_t = t \pm N \cdot K \cdot \sigma_{\bar{x}}$
Minimum Necessary Sample Size			$n = \frac{NZ^2 \cdot \sigma^2}{E^2 N + Z \hat{\sigma}^2}$	$n = \frac{NZ^2 \cdot \hat{P}(1-\hat{P})}{E^2 N + Z \cdot \hat{P}(1-\hat{P})}$

BEST COPY AVAILABLE

APPENDIX B: SAMPLING SYMBOLS AND FORMULAS (Continued)
(For Simple Random Sampling Without Replacement)

	Population		Sample	
	Continuous Variable	Categorical Variable	Continuous Variable	Categorical Variable
Minimum Necessary Sample Size (N unknown or very large)			$n = \frac{Z^2 \cdot \theta^2}{E^2}$	$n = \frac{Z^2 \cdot \hat{P}(1 - \hat{P})}{E^2}$

BEST COPY AVAILABLE

APPENDIX C

BIBLIOGRAPHY

There are a large number of books and articles on sampling statistics. For those wishing to take the next step beyond the materials presented in this manual, Mandenhall (1976), Storim (1960), and Sanders (1967) offer good treatments of basic sampling on an elementary mathematical level. Stuart (1962) covers the basic concepts of sampling in a presentation with very little mathematics. An excellent short presentation of stratified and cluster samples is contained in Lazewitz (1968). On an intermediate mathematical level, Jessen (1978) is a very useful reference. Hansen (1953), Cochran (1967), and Kish (1965) are valuable general books on statistical sampling which have become standard references. Volume II of Hansen contains the statistical derivations of most common formulas. For treatments of statistical sampling specifically related to the needs of auditors, see Arkin (1963) and Deming (1960).

BIBLIOGRAPHY

Anderson, R. and Teitebaun, A.D. "Dollar-unit Sampling" in Canadian Chartered Accountant, April 1973.

Arkin, H. 1963. Handbook of Sampling for Auditing and Accounting. McGraw-Hill, New York.

Cochran, W.G. 1967. Sampling Techniques. 3rd ed. Wiley, New York.

Deming, W.E. 1960. Sample Design in Business Research. Wiley, New York.

Hansen, M.H.; Kurwitz, W.N.; and Madow, W.G. 1953. Sample Survey Methods and Theory. Wiley, New York.

Jessen, R.J. 1978. Statistical Survey Techniques. Wiley, New York.

Kish, L. 1965. Survey Sampling. Wiley, New York.

Lazewitz, B. 1968. "Sampling Theory and Procedures" in H. Bladock Methodology in Social Research. 278-328. McGraw-Hill, New York.

Mendenhall, W.; Ott, L. and Sceaffer, R. 1976. Survey Sampling. Duxbury Press, Belmong, California.

Sanders, D.; Murph, A.F.; and Eng R.J. 1967. Statistics: A Fresh Approach.

Snedecor, G.W.; and Cochran, W.G. 1980. Statistical Methods. 7th ed. Iowa State University Press, Ames, Iowa.

Storim, M.J. 1960. Sampling in a Nutshell. Simon and Schuster, New York.

Stuart, A. 1962. Basic Ideas of Scientific Sampling. Griffin, London.

C2

INDEX

(Primary Reference or Definition)

Attributes	A6
Bias	11
Central Limit Theorem	A14
Cluster Sampling	22
Confidence Interval	13-14, A15-A16, A20-A21
Confidence Level	13-14
Confidence Limits	13-14
Discovery Sampling	29-30
Distribution	A6-A8
Dollar-Unit Sampling	26-27
Error, Sampling	10-13
Exploratory Sampling	29-30
Frequencies	A6-A7
Interval Estimation	13-14
Interval Sampling	23
Judgmental Sampling	5-8
Mean	A3
Mean Square Error	A13
Minimum Necessary Sample Size	A20-A22
Multi-stage Sampling	31
Multi-use Samples	53
Normal Distribution	A8
Opportunity Sampling	32
Parameter	10
Point Estimation	13
Population	10
Precision	11

INDEX (Continued)

Proportion	A6
Quota Sampling	33
Random Number Table	18
Replacement	20
Sample Size	41-45
Sampling	
Advantages	3-4
Disadvantages	4-5
Types of	15-34
Sampling Error	10-13
Sampling Unit	10
Sequential Sampling	28
Simple Random Sampling	16-20
Small Samples	A17-A18
Standard Deviation	
Standard Error of the Mean	A13
Statistics	10
Stop or Go Sampling	28
Stratified Sampling	21
Systematic Sampling	23
Tchebyshef's Theorem	A18
Variance	36-37, 47, A4-A5
Z Values	A8

☆U.S. GOVERNMENT PRINTING OFFICE: 1984 421 054 4243